Linear mixed-effects model for longitudinal complex data with diversified characteristics

Zhichao Wang, Huiwen Wang, Shanshan Wang, Shan Lu, Gilbert Saporta

Journal of Management Science and Engineering 5 (2020) 105–124

Contents lists available at ScienceDirect

Journal of Management Science and Engineering

journal homepage: www.keaipublishing.com/en/journals/journal-of-management-science-and-engineering/

Linear mixed-effects model for longitudinal complex data with diversified characteristics

Zhichao Wang [a], Huiwen Wang [a, b], Shanshan Wang [a, b, *], Shan Lu [c], Gilbert Saporta [d]

[a] School of Economics and Management, Beihang University, Beijing, China
[b] Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing, China
[c] School of Statistics and Mathematics, Central University of Finance and Economics, Beijing, China
[d] Applied Statistics, Conservatoire National des Arts et Metiers, Paris, 75141, France

ARTICLE INFO

Article history:
Available online 14 November 2019

Keywords:
Longitudinal complex data
Linear mixed-effects model
Compositional data analysis
Functional data analysis
Chinese stock market
Online investors' sentiment

ABSTRACT

The increasing richness of data encourages a comprehensive understanding of economic and financial activities, where variables of interest may include not only scalar (point-like) indicators, but also functional (curve-like) and compositional (pie-like) ones. In many research topics, the variables are also chronologically collected across individuals, which falls into the paradigm of longitudinal analysis. The complicated nature of data, however, increases the difficulty of modeling these variables under the classic longitudinal framework. In this study, we investigate the linear mixed-effects model (LMM) for such complex data. Different types of variables are first consistently represented using the corresponding basis expansions so that the classic LMM can then be conducted on them, which generalizes the theoretical framework of LMM to complex data analysis. A number of simulation studies indicate the feasibility and effectiveness of the proposed model. We further illustrate its practical utility in a real data study on Chinese stock market and show that the proposed method can enhance the performance and interpretability of the regression for complex data with diversified characteristics.

1. Introduction

The development of sensors, information storage and data mining makes it possible to collect data from a large number of sources with different characteristics, such as the familiar single points, curves and pie charts. Multiple types of data, referred to as *complex data*, greatly enlarge the traditional category of variables and provide researchers with an opportunity to understand the behavior of activities more comprehensively than ever before. For example, the public online sentiments measured from social media and intraday stock returns are improving the accuracy and interpretation of predicting trends in Chinese stock market (Wang et al., 2019a), which will be revisited in the real data analysis. Here, the return series are considered as the continuous functions from opening to closing and the investors' sentiments are measured as the compositions constituted by five types of specific sentiments. How to jointly analyze these emerging indicators with functional and compositional features and efficiently conduct models for complex data, such as regression considered in this paper, has been a serious challenge for economical modeling.

* Corresponding author. School of Economics and Management, Beihang University, Beijing, China.
E-mail addresses: sswang@buaa.edu.cn (S. Wang), shan.lu@cufe.edu.cn (S. Lu).

https://doi.org/10.1016/j.jmse.2019.11.001

Complex data analysis involves various types of data, ranging from the classic nominal, ordinal and ratio scalar variables to curve- and pie-like functional and compositional variables and even the text data. In this study, we focus on two novel types, namely the functional and compositional data. Specifically, in functional data analysis (FDA), a data unit is assumed to be a square-integrable function determined by its observations at various times (Ramsay, 1982). The intrinsic property of functions, namely the infinite dimension, causes great difficulty for functional modeling in both theory and implementation. To describe functions over a bounded closed set (e.g., an interval), some equivalent representations such as the basis function expansion and reproducing kernel methods are thus necessary (Härdle et al., 2012). On the contrary, compositional data analysis (CDA) discusses the intrinsic structure of a whole, such as the proportions or percentages that carry only the relative information (Aitchison, 1982, 1986). The defining features of compositions include the strictly positive and constant-sum constraints of all the components inside (e.g., 1 for proportions and 100 for percentages), which would be problematic for most traditional statistical approaches. To eliminate these strong constraints, a family of logratio transformations have been proposed, such as the additive logratio, centered logratio (Aitchison, 1986) and isometric logratio (Egozcue et al., 2003) transformations. For further details on FDA and CDA, see Ramsay and Silverman (2005, 2007), and Pawlowsky-Glahn et al. (2015) and Filzmoser et al. (2018), respectively.

Numerous works have investigated the regression for the functional/compositional covariate against a scaler response. Ramsay and Silverman (2005, 2007) systematically proposed the theoretical framework of functional linear regression, and Müller and Stadtmüller (2005) then expanded it to the generalized linear case. More recent studies of functional regression employ the additive model (Fan et al., 2015), mixture of linear models (Wang et al., 2016) and truncated linear model (Hall & Hooker, 2016). Meanwhile, Aitchison and Bacon-Shone (1984) initially proposed the linear regression for compositional covariates, Marzio et al. (2015) presented the kernel-based compositional regression, and Bruno, Greco, and Ventrucci (2014, 2016) investigated the spatio-temporal model and another nonparametric regression with Bayesian P-splines, respectively.

The aforementioned approaches focus on a specific type of complex data, and relatively few studies examine the multi-type situation. Wang et al. (2015) preliminarily developed the linear regression for multiple types of complex data. Wang et al. (2019a) then extended the computational framework to generalized linear regression including scalar, functional and compositional covariates. Their methods are based on the independent and identically distributed (IID) assumption on errors, namely that all the samples of complex data, regarded as cross-sectional data, are assumed to be independently collected from an identical population. However, this IID assumption is problematic in some cases; these complex data may also show typical longitudinal features, which we call *longitudinal complex data*. For example, as discussed in Section 5, the closing prices of Chinese stock market were collected from hundreds of stocks over several months. The price-limiting mechanism of the market results in a high correlation among the observations of the same stock. Directly applying existing statistical models to complex data without incorporating the correlation structure might yield biased and confusing results. Similar problems have been discussed separately for FDA (Chen & Cao, 2017; Gertheiss et al., 2013; Goldsmith et al., 2012) and CDA (Qiu et al., 2010; Wang et al., 2019b; Zhang et al., 2009), but few of the related solutions show the potential to integrate multiple types of complex data together. Thus, developing a unified framework of modeling longitudinal complex data with diversified characteristics is worthy of study.

As a fundamental longitudinal technique, the linear mixed-effects model (LMM) proposed by Laird and Ware (1982) has been extended to numerous applications (Fitzmaurice et al., 2011; Hsiao, 2014), but most studies are concerned only with scalar variables. In this study, we investigate LMM for complex data with diversified characteristics (CompLMM hereafter) to deal with longitudinal features. Specifically, we assume that the data are collected from N individuals along with n_i measurements for the i-th individual ($i = 1, 2, \ldots, N$); then, CompLMM is formulated as

$$y_{ij} = \sum_{k=1}^{p_x} x_{ijk}\alpha_k + \sum_{k=1}^{q_z} z_{ijk}a_{ik} + \sum_{k=1}^{p_\mu}\int \mu_{ijk}\beta_k + \sum_{k=1}^{q_\nu}\int \nu_{ijk}b_{ik} + \sum_{k=1}^{p_c}\left(\boldsymbol{c}_{ijk}, \boldsymbol{\gamma}_k\right)_a + \sum_{k=1}^{q_w}\left(\boldsymbol{w}_{ijk}, \boldsymbol{r}_{ik}\right)_a + \varepsilon_{ij} \tag{1}$$

for the j-th sample ($j = 1, 2, \ldots, n_i$), where $(\cdot, \cdot)_a$ denotes the Aitchison inner product in CDA to be introduced in Section 2.2. Take the stock market for example, in Model (1), the scalar response y_{ij} may refer to the closing price; the scalar, functional and compositional covariates, namely x_{ijk} (z_{ijk}), μ_{ijk} (ν_{ijk}) and $\boldsymbol{c}_{ijk}$ ($\boldsymbol{w}_{ijk}$), may refer to the classic indicators like the daily volume, curve-like intraday stock pricing, and pie-like proportions of investors' sentiments, with the dimensions of p_x (q_z), p_μ (q_ν) and p_c (q_w), respectively; α_k (a_{ik}), β_k (b_{ik}) and $\boldsymbol{\gamma}_k$ ($\boldsymbol{r}_{ik}$) denote the regression coefficients with the corresponding characteristics; and ε_{ij} is the random scalar error. Moreover, the functional covariates may also include yearly/monthly micro-indexes in economics and high-frequency transaction data in finance, and the compositional ones can be used for portfolio weights.

In the paradigm of LMM, the terms containing α_k, β_k and $\boldsymbol{\gamma}_k$ in Model (1) comprise the fixed effects shared by all individuals, whereas those containing a_{ik}, ν_{ik} and $\boldsymbol{r}_{ik}$ comprise the random effects unique to the specific individual. Particularly, when there are no random effects, Model (1) is categorized as the computational framework of complex data in Wang et al. (2019a) and reduces to the IID-based linear model (Wang et al., 2015), called CompLM, as

$$y_{ij} = \sum_{k=1}^{p_x} x_{ijk}\alpha_k + \sum_{k=1}^{p_\mu}\int \mu_{ijk}\beta_k + \sum_{k=1}^{p_c}\left(\boldsymbol{c}_{ijk}, \boldsymbol{\gamma}_k\right)_a + \varepsilon_{ij}.$$

Compared with CompLM, the introduction of the random effects in CompLMM makes it possible to capture the subject-specific information of each individual on the basis of the common regression characteristics in the population. From the regression error perspective, it extracts the variability caused by different individuals, namely $\sum_{k=1}^{q_z} z_{ijk}a_{ik} + \sum_{k=1}^{q_\nu} \int \nu_{ijk}b_{ik} + \sum_{k=1}^{q_w} (\boldsymbol{w}_{ijk}, \boldsymbol{r}_{ik})_a$, from the OLS-estimated errors. That is, ε_{ij} is further decomposed into the random effects explained by individuals and the normally distributed noise. Therefore, the proposed CompLMM distinguishes the between- and within-subject variability of the responses, which improves the performance of the linear regression on longitudinal complex data.

In this study, we aim at estimating the parameters for Model (1). Following the diversified characteristics of covariates, both fixed and random effects show the corresponding characteristics. This brings the theoretical challenge in understanding the relationship between the scalar error and multiple types of fixed/random effects. To consistently represent the complex data with diversified characteristics, we first transform the functions and compositions using the related basis expansions such that they are described equivalently or approximately equivalently to numeric coordinates. Specifically, the principle component analysis (PCA) for functional data is introduced to generate a group of orthonormal basis functions according to the observations. The transformed data are available to conduct LMM and obtain the intermediate results that can be reconstructed to match the original features. Moreover, we investigate the statistical inferences on the coefficients of the fixed effects. The necessary theoretical properties for the proposed longitudinal framework are developed accordingly. To further measure the variability of different types of variables across individuals, we adopt the point-wise variance function and total variance for a random function and composition, respectively. The proposed CompLMM improves the regression of complex data with diversified characteristics and enhances its interpretability, which may provide an instructive unified framework for modeling longitudinal complex data.

The remainder of this paper is organized as follows. In Section 2, we review some fundamental knowledge on FDA and CDA. In Section 3, we investigate CompLMM and propose its computational algorithm. A series of simulation studies are then conducted to assess the performance of the proposed method, with the results presented in Section 4. Section 5 reports a real data study on Chinese stock market to illustrate the effectiveness of the proposed method. Finally, some conclusions and prospects are given in Section 6.

2. Preliminaries

We briefly introduce the basic ideas and mathematical techniques for FDA and CDA, including the basis function expansion for functional data, and the Aitchison geometry, centered logratio (clr) and isometric logratio (ilr) transformations for compositional data. These provide the theoretical and computational foundation for the proposed method. For simplicity, we use commas and semicolons in the matrix expressions to indicate that the adjacent blocks in a matrix are organized by column and row, respectively.

2.1. FDA

In FDA, a series of discrete data are considered to be collected from a potential single entity (i.e., the function) over a continuous index such as time and space.

The basis function expansion is one of the most practical methods for describing the continuous characteristics of a function (Ramsay & Silverman, 2005). That is, a function is expressed as a linear combination of a given group of basis functions, which can be estimated using the ordinary least squares (OLS), penalized OLS or regularized principal component method (Hall & Horowitz, 2007). Two popular basis function systems in FDA are the Fourier and B-spline basis systems. A series of Fourier basis functions with different periods are more appropriate for expanding periodic functions and describing the periodic features. By contrast, B-spline functions are used more frequently for those non-periodic functions since they can be flexibly identified by many parameters such as the number and locations of the inner knots. Without loss of generality, we adopt the B-spline basis functions and simple OLS-based estimation.

Specifically, given a group of basis functions $\{\phi_j\}_{j=1}^{\infty}$ over the interval $\mathcal{I}$, any square-integrable function, say $\mu \in \mathscr{L}^2$, can be formulated as $\mu = \sum_j u_j\phi_j$ with an infinite series of expansion coefficients u_j's. When n samples, say o_i at time $t_i \in \mathscr{I}$ ($i = 1, 2, \ldots, n$), are observed from μ, they are assumed to follow $o_i = \mu(t_i) + \varepsilon_i$ with the white noise ε_i. Then, the expansion of μ leads to $\boldsymbol{o} = \boldsymbol{\Phi u} + \varepsilon$, where $\boldsymbol{o} = (o_1, o_2, \ldots, o_n)'$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ in $\mathbb{R}^n$, $\boldsymbol{\Phi} = (\phi(t_1), \phi(t_2), \ldots, \phi(t_n))'$ with $\phi(t_i) = (\phi_1(t_i), \phi_2(t_i), \ldots, \phi_K(t_i))'$, and $\boldsymbol{u} = (u_1, u_2, \ldots, u_K)'$. In practice, the number of basis functions and related expansion coefficients K are restricted to be smaller than n because of the finite size of observations. Thus, the OLS estimation yields the truncated expansion coefficients of μ, namely

$$\boldsymbol{u} = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{o}. \tag{2}$$

Using the truncated expansion coefficients, the curve of μ can be drawn numerically in a point-wise manner as

$$\mu(t) = \sum_{j=1}^{K} u_j\phi_j(t) = \boldsymbol{u}'\phi(t) \quad (t \in \mathscr{I}), \tag{3}$$

where μ is determined by $\boldsymbol{u}$ given the basis functions $\phi = (\phi_1, \phi_2, \ldots, \phi_K)'$; the variance function for μ, denoted by $\mathscr{K}_\mu$, can also be formulated as

$$\mathscr{K}_\mu(t) = \mathrm{Var}(\mu(t)) = \phi'(t)\mathrm{Var}(\boldsymbol{u})\phi(t) \quad (t \in \mathscr{I}), \tag{4}$$

where $\mathrm{Var}(\cdot)$ denotes the covariance matrix of a random vector. Moreover, the integral of the product of two functions, say μ and β with the expansion coefficients $\boldsymbol{u}$ and $\boldsymbol{\lambda}$, can be written as

$$\int \mu(t)\beta(t)\mathrm{d}t = \boldsymbol{u}'\boldsymbol{W}\boldsymbol{\lambda} \quad \text{with} \quad \boldsymbol{W} = \int \phi(t)\phi'(t)\mathrm{d}t. \tag{5}$$

The similar expression for the squared distance between two functions follow

$$d^2_{\mathcal{L}^2}(\beta, \widehat{\beta}) = \int (\beta(t) - \widehat{\beta}(t))^2\mathrm{d}t = (\boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}})'\boldsymbol{W}(\boldsymbol{\lambda} - \widehat{\boldsymbol{\lambda}}) \tag{6}$$

for $\widehat{\beta} \in \mathcal{L}^2$ with its expansion coefficients $\widehat{\boldsymbol{\lambda}}$ under ϕ.

From the aforementioned properties, the truncated expansion coefficients greatly concentrate the main features of the original infinite-dimensional functional data. However, the elimination on the other basis functions can unavoidably lead to the extra approximation bias in the error. Intuitively, a larger group of basis functions would provide a more detailed description on the observed functional data; therefore, the approximately bias could get reasonably close to zero if the numbers of the observations and basis functions are large enough. For the practical purpose, we may empirically choose the basis function system or perform the dimension reduction procedure to realize the efficient smoothing for functions with little loss of information. Finally, we conclude that the basis function expansion for functional data makes it possible to approximately equivalently represent the infinite-dimensional functions as relatively few numeric variables.

2.2. CDA

In CDA, the focus is on the relative magnitude of all components within a multivariate vector, rather than the absolute value. Mathematically, any composition with D inner parts can be expressed as a vector $\boldsymbol{c} = [c_1, c_2, \ldots, c_D]'$ with the positive and featuring constant-sum constraints, namely

$$0 < c_i < 1 \quad (i = 1, 2, \ldots, D) \quad \text{and} \quad \sum_{i=1}^{D} c_i = 1.$$

All such D-part compositions consist of the D-dimensional simplex space denoted by S^D. To highlight the element in the simplex space, we use the square bracket for compositional data throughout this paper.

Due to these constraints, the classic Euclidean algebraic system is not appropriate for the simplex space. For example, the components of the sum of two compositions may exceed the range between 0 and 1. To construct the linear space for compositional data, Aitchison (1982) developed the Aitchison geometry. Two main operations in the Aitchison geometry are the perturbation and powering, and for $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \ldots, \gamma_D]' \in S^D$ and $\kappa \in \mathbb{R}$, they are respectively

$$\boldsymbol{c} \oplus \boldsymbol{\gamma} = \mathcal{C}(c_1\gamma_1, c_2\gamma_2, \ldots, c_D\gamma_D) \quad \text{and} \quad \kappa \odot \boldsymbol{c} = \mathscr{C}\left(c_1^\kappa, c_2^\kappa, \ldots, c_D^\kappa\right),$$

where $\mathcal{C}(\cdot)$ denotes the closure operation that scales a vector with all positive components proportionally such that it conforms to the constant-sum constraint. The zero element is $\boldsymbol{n}^D = \mathcal{C}(1_{\boldsymbol{D}})$ with $1_D = (1, 1, \ldots, 1)' \in \mathbb{R}^D$. The related inner product for two compositions are then defined as

$$(\boldsymbol{c}, \boldsymbol{\gamma})_a = \sum_{i=1}^{D} \log\frac{c_i}{\mathrm{g}_m(\boldsymbol{c})} \log\frac{\gamma_i}{\mathrm{g}_m(\boldsymbol{\gamma})}, \tag{7}$$

where $\mathrm{g}_m(\cdot)$ denotes the geometric mean of all parts in a composition.

The Aitchison inner product enjoys nice properties for the simplex space but lacks an intuitive expression. To simplify the formula, Aitchison (1986) proposed the clr transformation, namely

$$\mathrm{clr}(\boldsymbol{c}) = (\mathrm{clr}_1(\boldsymbol{c}), \mathrm{clr}_2(\boldsymbol{c}), \ldots, \mathrm{clr}_D(\boldsymbol{c}))' = \left(\log\frac{c_1}{\mathrm{g}_m(\boldsymbol{c})}, \log\frac{c_2}{\mathrm{g}_m(\boldsymbol{c})}, \ldots, \log\frac{c_D}{\mathrm{g}_m(\boldsymbol{c})}\right)',$$

Such that (7) can be rewritten in the Euclidean form as $(\boldsymbol{c}, \boldsymbol{\gamma})_a = \mathrm{clr}(\boldsymbol{c})'\mathrm{clr}(\boldsymbol{\gamma})$. By comparing each component with the geometric mean, the clr-transformed vector describes the relative magnitude of the corresponding part in the original

composition. Specifically, $\mathrm{clr}_i(\boldsymbol{c}) > 0$ for some i ($i = 1, 2, \ldots, D$) indicates that the proportion of the i-th part in $\boldsymbol{c}$ is above the average level of all parts, and vise versa.

The clr transformation simplifies the expression of Aitchison geometry but remains constrained still, namely $\sum_{i=1}^{D} \mathrm{ilr}_i(\boldsymbol{c}) = 1$. To represent compositional data with no constraints simultaneously, Egozcue et al. (2003) further proposed the ilr transformation via the simplicial orthonormal basis. In this study, we follow Egozcue and Pawlowsky-Glahn (2005) and transform the composition to the specified ilr coordinates as $\mathrm{ilr}(\boldsymbol{c}) = \boldsymbol{c}^* = (c_1^*, c_2^*, \ldots, c_{D-1}^*)'$, where

$$c_i^* = \frac{1}{\sqrt{(D-i+1)(D-i)}} \sum_{j=1}^{D-i} \log c_j - \sqrt{\frac{D-i}{D-i+1}}\, \log c_{D-i+1} \quad (i = 1, 2, \ldots, D-1). \tag{8}$$

These coordinates contain all the information on $\boldsymbol{c}$, and they can reconstruct the original composition. That is, $\boldsymbol{c} = \mathrm{ilr}^{-1}(\boldsymbol{c}^*) = \mathcal{C}(\exp(\boldsymbol{\omega}))$, where $\exp(\boldsymbol{\omega}) = (\exp\omega_1, \exp\omega_2, \ldots, \exp\omega_D)'$,

$$\omega_i = \sum_{j=0}^{D-i} \frac{c_j^*}{\sqrt{(D-j+1)(D-j)}} - \sqrt{\frac{i-1}{i}} c_{D-i+1}^* \quad (i = 1, 2, \ldots, D) \tag{9}$$

and $c_0^* = c_D^* = 0$. Using the contrast matrix, denoted by $\boldsymbol{\Psi} \in \mathbb{R}^{(D-1)\times D}$, the ilr transformation and its inverse can be respectively expressed as $\mathrm{ilr}(\boldsymbol{c}) = \boldsymbol{\Psi} \log(\boldsymbol{c})$ and $\mathrm{ilr}^{-1}(\boldsymbol{c}^*) = \mathscr{C}(\exp(\boldsymbol{\Psi}'\boldsymbol{c}^*))$, where $\log(\boldsymbol{c}) = (\log c_1, \log c_2, \ldots, \log c_D)'$. Specifically, $\boldsymbol{\Psi}$ associated with (8) and (9) is constituted by the elements ψ_{ij} as $\psi_{ij} = \sqrt{\frac{D-i}{D-i+1}}\rho_{ij}$ for $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$, where $\rho_{ij} = (D-i)^{-1}$ when $j < D-i+1$, $\rho_{ij} = -1$ when $j = D-i+1$, and $\rho_{ij} = 0$ otherwise.

As an isometry between the simplex and Euclidean spaces, the ilr transformation facilitates the computation of the Aitchison geometry. Using the ilr coordinates, the Aitchison inner product in (7) can then be expressed in the familiar Euclidean form as

$$(\boldsymbol{c}, \boldsymbol{\gamma})_a = \mathrm{ilr}(\boldsymbol{c})'\mathrm{ilr}(\boldsymbol{\gamma}) = (\boldsymbol{c}^*)'\boldsymbol{\gamma}^*. \tag{10}$$

The related norm and distance then follow respectively as

$$\|\boldsymbol{\gamma}\|_a^2 = (\boldsymbol{\gamma}, \boldsymbol{\gamma})_a = (\boldsymbol{\gamma}^*)'\boldsymbol{\gamma}^* \quad \text{and} \quad d_a^2(\boldsymbol{\gamma}, \widehat{\boldsymbol{\gamma}}) = (\boldsymbol{\gamma}^* - \widehat{\boldsymbol{\gamma}}^*)'(\boldsymbol{\gamma}^* - \widehat{\boldsymbol{\gamma}}^*)$$

with $\widehat{\boldsymbol{\gamma}} \in S^D$ and its ilr coordinates $\widehat{\boldsymbol{\gamma}}^*$. Since $\boldsymbol{\Psi}'\boldsymbol{\Psi}$ is identically equal to $\boldsymbol{I}_D - 1_D 1_D'/D$, where $\boldsymbol{I}_D$ denotes the D-dimensional identical matrix (Pawlowsky-Glahn et al., 2015), the specified ilr transformation here does not affect those results above. Moreover, taking values in $\mathbb{R}^{D-1}$ freely, $\boldsymbol{c}^*$ and $\boldsymbol{\gamma}^*$ become the representations of $\boldsymbol{c}$ and $\boldsymbol{\gamma}$ with no constraints involved. From these properties of the ilr transformation, we can substitute the ilr coordinates for the compositional covariates in most statistical models.

Finally, the moments of a random composition are different from those of a scalar random variable due to the lack of the multiplication operation on the simplex space. Pawlowsky-Glahn et al. (2015) defined the total variance and center of a random composition as

$$\mathrm{totVar}[\boldsymbol{c}] = \min_{\boldsymbol{u} \in S^D} \{\mathrm{Var}[\boldsymbol{c}; \boldsymbol{u}]\} \quad \text{and} \quad \mathrm{cen}[\boldsymbol{c}] = \arg\min_{\boldsymbol{u} \in S^D} \{\mathrm{Var}[\boldsymbol{c}; \boldsymbol{u}]\},$$

Respectively, where $\mathrm{Var}[\boldsymbol{c}; \boldsymbol{u}] = \mathbb{E}[d_a^2(\boldsymbol{x}, \boldsymbol{u})]$ and $\mathbb{E}[\cdot]$ denotes the expectation of a random variable. The center and total variance of a random composition serve as the mean and variance of a scalar random variable, and they can be respectively formulated as

$$\mathrm{cen}[\boldsymbol{c}] = \mathrm{ilr}^{-1}(\mathbb{E}[\mathrm{ilr}(\boldsymbol{c})]) \quad \text{and} \quad \mathrm{totVar}[\boldsymbol{c}] = \sum_{i=1}^{D-1} \mathrm{Var}(c_i^*)$$

via the ilr transformation.

3. LMM for complex data

In this section, we investigate LMM for longitudinal complex data with diversified characteristics. The approaches to aggregating multiple types of complex data, along with their properties, and some issues in implementation are also discussed.

3.1. Model

To consistently represent multiple types of complex data, we apply the B-spline function expansion and ilr transformation to Model (1). Thus, the model can be formulated using (5) and (10) as

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\alpha} + \boldsymbol{z}_{ij}'\boldsymbol{a}_i + \sum_{k=1}^{p_\mu} \boldsymbol{u}_{ijk}'\boldsymbol{W}\boldsymbol{\lambda}_k + \sum_{k=1}^{q_\nu} \boldsymbol{v}_{ijk}'\boldsymbol{W}\boldsymbol{\theta}_{ik} + \sum_{k=1}^{p_c} \left(\boldsymbol{c}_{ijk}^*\right)' \boldsymbol{\gamma}_k^* + \sum_{k=1}^{q_w} \left(\boldsymbol{w}_{ijk}^*\right)' \boldsymbol{r}_{ik}^* + \varepsilon_{ij},$$

where $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ij,p_x})'$ and $\boldsymbol{z}_{ij} = (z_{ij1}, z_{ij2}, \ldots, z_{ij,q_z})'$ along with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{p_x})'$ and $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \ldots, a_{i,q_z})'$; $\boldsymbol{u}_{ijk}$, $\boldsymbol{v}_{ijk}$, $\boldsymbol{\lambda}_k$ and $\boldsymbol{\theta}_{ik}$ denote the expansion coefficients for μ_{ijk}, ν_{ijk}, β_k and b_{ik}, respectively; and $\boldsymbol{c}_{ijk}^*$, $\boldsymbol{\gamma}_k^*$, $\boldsymbol{w}_{ijk}^*$ and $\boldsymbol{r}_{ik}^*$ denote the ilr coordinates of the related compositions. To simplify, we further reformulate it as

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\alpha} + \boldsymbol{z}_{ij}'\boldsymbol{a}_i + \boldsymbol{u}_{ij}'\boldsymbol{W}_{p_\mu}\boldsymbol{\lambda} + \boldsymbol{v}_{ij}'\boldsymbol{W}_{q_\nu}\boldsymbol{\theta}_i + \left(\boldsymbol{c}_{ij}^*\right)'\boldsymbol{\gamma}^* + \left(\boldsymbol{w}_{ij}^*\right)'\boldsymbol{r}_i^* + \varepsilon_{ij}, \tag{11}$$

where $\boldsymbol{u}_{ij} = (\boldsymbol{u}_{ij1}; \boldsymbol{u}_{ij2}; \ldots; \boldsymbol{u}_{ij,p_\mu})$ and $\boldsymbol{v}_{ij} = (\boldsymbol{v}_{ij1}; \boldsymbol{v}_{ij2}; \ldots; \boldsymbol{v}_{ij,q_\nu})$ along with $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1; \boldsymbol{\lambda}_2; \ldots; \boldsymbol{\lambda}_{p_\mu})$ and $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}; \boldsymbol{\theta}_{i2}; \ldots; \boldsymbol{\theta}_{i,q_\mu})$; $\boldsymbol{W}_{p_\mu}$ ($\boldsymbol{W}_{q_\nu}$) denotes the blocked diagonal matrix consisting of p_μ (q_ν) matrices $\boldsymbol{W}$; and $\boldsymbol{c}_{ij}^* = (\boldsymbol{c}_{ij1}^*; \boldsymbol{c}_{ij2}^*; \ldots; \boldsymbol{c}_{ij,p_c}^*)$ and $\boldsymbol{w}_{ij}^* = (\boldsymbol{w}_{ij1}^*; \boldsymbol{w}_{ij2}^*; \ldots; \boldsymbol{w}_{ij,q_w}^*)$ along with $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}_1^*; \boldsymbol{\gamma}_2^*; \ldots; \boldsymbol{\gamma}_{p_c}^*)$ and $\boldsymbol{r}_i^* = (\boldsymbol{r}_{i1}^*; \boldsymbol{r}_{i2}^*; \ldots; \boldsymbol{r}_{i,q_w}^*)$, respectively. Note that the dimensions of related expansion coefficients (e.g., $\boldsymbol{u}_{ijk}$ for $k = 1, 2, \ldots, p_\mu$) may differ in functional variables as we can use different basis functions with different numbers (K) and even types to represent the functional data to better describe their potentially diversified features.

Jointly considering all samples from the same individual, say the i-th one, we pile up the related n_i samples by row. Finally, Model (11) can be rewritten as

$$\boldsymbol{y}_i = \boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{u}_i\boldsymbol{W}_{p_\mu}\boldsymbol{\lambda} + \boldsymbol{c}_i^*\boldsymbol{\gamma}^* + \boldsymbol{z}_i\boldsymbol{a}_i + \boldsymbol{v}_i\boldsymbol{W}_{q_\nu}\boldsymbol{\theta}_i + \boldsymbol{w}_i^*\boldsymbol{r}_i^* + \boldsymbol{\varepsilon}_i \quad (i = 1, 2, \ldots, N), \tag{12}$$

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,n_i})'$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{i,n_i})'$. Specifically, in Model (12), the fixed effects involve $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{i,n_i})'$, $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \ldots, \boldsymbol{u}_{i,n_i})'$ and $\boldsymbol{c}_i^* = (\boldsymbol{c}_{i1}^*, \boldsymbol{c}_{i2}^*, \ldots, \boldsymbol{c}_{i,n_i}^*)'$, and the random effects involve $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \ldots, \boldsymbol{z}_{i,n_i})'$, $\boldsymbol{v}_i = (\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \ldots, \boldsymbol{v}_{i,n_i})'$ and $\boldsymbol{w}_i^* = (\boldsymbol{w}_{i1}^*, \boldsymbol{w}_{i2}^*, \ldots, \boldsymbol{w}_{i,n_i}^*)'$. To coincide with the paradigm of LMM in Model (12), the total coefficients for the fixed and random effects refer to $\boldsymbol{\varpi} = (\boldsymbol{\alpha}; \boldsymbol{\lambda}; \boldsymbol{\gamma}^*)$ and $\boldsymbol{\pi}_i = (\boldsymbol{a}_i; \boldsymbol{\theta}_i; \boldsymbol{r}_i^*)$, with dimensions p and q, respectively. Moreover, $\boldsymbol{\varepsilon}_i$ is assumed to be normally distributed in $\mathbb{R}^{n_i}$, namely $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(0_{n_i}, \sigma^2\boldsymbol{I}_{n_i})$, where 0_{n_i} is constituted by 0 in $\mathbb{R}^{n_i}$, and $\boldsymbol{\pi}_i$ is assumed to be independent of $\boldsymbol{\varepsilon}_i$ and normally distributed as $\boldsymbol{\pi}_i \sim \mathcal{N}(0_q, \boldsymbol{G})$, where $\boldsymbol{G}$ is positive definite and constant for all the individuals. Thus, the parameters to be estimated in Model (12) include $\boldsymbol{\Theta} = \{\boldsymbol{\varpi}, \boldsymbol{G}, \sigma^2\}$.

Under the aforementioned assumptions, the estimate of $\boldsymbol{\Theta}$, denoted by $\widehat{\boldsymbol{\Theta}} = \{\widehat{\boldsymbol{\varpi}}, \widehat{\boldsymbol{G}}, \widehat{\sigma}^2\}$, can be derived using the expectation maximum (EM) algorithm (Laird & Ware, 1982). Then, the fitted response, for example, $\widehat{y}_{ij}$ in Model (11), can be expressed as

$$\widehat{y}_{ij} = \boldsymbol{x}_{ij}'\widehat{\boldsymbol{\alpha}} + \boldsymbol{z}_{ij}'\widehat{\boldsymbol{a}}_i + \boldsymbol{u}_{ij}'\boldsymbol{W}_{p_\mu}\widehat{\boldsymbol{\lambda}} + \boldsymbol{v}_{ij}'\boldsymbol{W}_{q_\nu}\widehat{\boldsymbol{\theta}}_i + \left(\boldsymbol{c}_{ij}^*\right)'\widehat{\boldsymbol{\gamma}}^* + \left(\boldsymbol{w}_{ij}^*\right)'\widehat{\boldsymbol{r}}_i^*,$$

or $\widehat{y}_{ij} = \boldsymbol{x}_{ij}'\widehat{\boldsymbol{\alpha}} + \boldsymbol{u}_{ij}'\boldsymbol{W}_{p_\mu}\widehat{\boldsymbol{\lambda}} + (\boldsymbol{c}_{ij}^*)'\widehat{\boldsymbol{\gamma}}^*$ for the reduced CompLM, where $\widehat{\boldsymbol{a}}_i$, $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{r}}_i^*$ can be obtained from $\widehat{\boldsymbol{\Theta}}$.

The assumptions on multiple types of complex data are based on the consistent numeric representations in the classic framework of LMM. Actually, the multivariate normal distribution for these original complex data is not the same as that for the classic multivariate case. Since different types of complex data involved, the relationship inside the random effects is not limited to the classic case for multivariate variables. Directly defining the joint distribution on the random effects with diversified characteristics would be difficult, as it may contain complicate covariance structures such as the correlation between a composition and a function. Before we develop the estimation procedure, we discuss the relationship between the original and transformed models, namely Models (1) and (12), in the following remarks.

Remark 1. The assumption of the independence of the coefficients for random effects and the error is important in the theory of LMM. We put this assumption on the represented numeric variables, namely $\boldsymbol{\theta}_i$ and $\boldsymbol{r}_i^*$, in Model (12), which also implies the independence of the original coefficients with diversified characteristics (e.g., b_{ik} and $\boldsymbol{r}_{ik}$) and the scalar error in Model (1). For example, the covariance matrix between b_{ik} with functional characteristics and ε_{ij} is defined in a point-wise manner, namely $\mathrm{Cov}(b_{ik}(t), \varepsilon_{ij})$ for any $t \in \mathcal{I}$. Under the given group of basis functions ϕ, the expectations of the related expansion coefficients are equal to zero; then, we have

$$\mathrm{Cov}\left(b_{ik}(t), \varepsilon_{ij}\right) = \boldsymbol{\theta}_{ik}{}'\mathbb{E}\left[\phi(t)\varepsilon_{ij}\right] = \phi'(t)\mathrm{Cov}\left(\boldsymbol{\theta}_{ik}, \varepsilon_{ij}\right). \tag{13}$$

From (13), the independence assumption on the expansion coefficients, namely $\mathrm{Cov}(\boldsymbol{\theta}_{ik}, \varepsilon_{ij}) = 0_K$, is sufficient to that on the original overall function. Meanwhile, the covariance between $\boldsymbol{r}_{ik}$ with the compositional characteristics and ε_{ij} is directly defined using the ilr coordinates, namely $\mathrm{Cov}(\boldsymbol{r}_{ik}^*, \varepsilon_{ij})$, and the independence assumption holds for any specified ilr transformation (Wang et al., 2019).

Remark 2. The most fundamental distribution, namely the multivariate normal distribution, for the consistent numeric representation of random effects is considered in this paper. The zero-mean assumption on the expansion coefficients for functional data and ilr coordinates for compositional data actually implies the related forms on the original random effects with diversified characteristics. Specifically, the center of $\boldsymbol{r}_{ik}$ will be the zero element in S^D since

$$\mathrm{cen}[\boldsymbol{r}_{ik}] = \mathrm{ilr}^{-1}\left(\mathbb{E}\left[\boldsymbol{r}_{ik}^*\right]\right) = \mathrm{ilr}^{-1}(0_D) = \boldsymbol{n}^D,$$

and the expectation of b_{ik} will also be the constant-zero function in a point-wise manner, namely

$$\mathbb{E}[b_{ik}(t)] = \phi'(t)\mathbb{E}[\boldsymbol{\theta}_{ik}] = 0$$

for any $t \in \mathcal{I}$.

Remark 3. Another issue for the theory of LMM follows the covariance matrix of the random effects coefficients. For functional variables, this refers to the covariance function, say $\mathscr{K}_{b_{ik},b_{ik'}}(s,t)$ for b_{ik} and $b_{ik'}$ with the expansion coefficients $\boldsymbol{\theta}_{ik}$ and $\boldsymbol{\theta}_{ik'}$ at s and t. Similar to (4), it can be formulated as

$$\mathcal{K}_{b_{ik},b_{ik'}}(s,t) = \mathrm{Cov}(b_{ik}(s), b_{ik'}(t)) = \phi'(s)\mathrm{Cov}(\boldsymbol{\theta}_{ik}, \boldsymbol{\theta}_{ik'})\phi(t).$$

Specifically, it reduces to $\mathcal{K}_{b_{ik}}(s,t) = \phi'(s)\mathrm{Var}(\boldsymbol{\theta}_{ik})\phi(t)$ when $k = k'$. For compositional variables, the covariance matrix can be naturally defined as $\mathrm{Cov}(\boldsymbol{r}_{ik}, \boldsymbol{r}_{ik'}) = \mathrm{Cov}(\boldsymbol{r}_{ik}^*, \boldsymbol{r}_{ik'}^*)$ for $\boldsymbol{r}_{ik}$ and $\boldsymbol{r}_{ik'}$ with the ilr coordinates $\boldsymbol{r}_{ik}$ and $\boldsymbol{r}_{ik'}^*$. If we adopt another distribution on the simplex, such as the Dirichlet one, the covariance structure for the original compositional data could be singular due to the constant-sum constraint of a composition. It would also be difficult to define the independence between a composition and the scalar error; by contrast, employing the ilr coordinates can explicitly avoid these troubles. For example, the covariance structure for the clr-transformed vector of $\boldsymbol{r}_{ik}$ can be easily derived as $\boldsymbol{\Phi}'\boldsymbol{\Sigma}\boldsymbol{\Phi}$. Next, the covariance matrix for the two types of variables can be defined consistently. For example, we express the covariance function for b_{ik} and $\boldsymbol{r}_{ik}$ at time t, denoted by $\mathcal{K}_{b_{ik},\boldsymbol{r}_{ik}}(t)$, as

$$\mathcal{K}_{b_{ik},\boldsymbol{r}_{ik}}(t) = \mathrm{Cov}\left(b_{ik}(t), \boldsymbol{r}_{ik}^*\right) = \phi'(t)\mathrm{Cov}\left(\boldsymbol{\theta}_{ik}, \boldsymbol{r}_{ik}^*\right).$$

In those cases, the different patterns inside the covariance matrix for the original model are described by the elements of $\boldsymbol{G}$, including $\mathrm{Cov}(\boldsymbol{\theta}_{ik}, \boldsymbol{\theta}_{ik'})$, $\mathrm{Var}(\boldsymbol{\theta}_{ik})$, $\mathrm{Cov}(\boldsymbol{r}_{ik}^*, \boldsymbol{r}_{ik'}^*)$ and $\mathrm{Cov}(\boldsymbol{\theta}_{ik}, \boldsymbol{r}_{ik}^*)$. Thus, $\boldsymbol{G}$ in the transformed model contains the covariance structure of the original complex data.

3.2. Parameter estimation

When there are no random effects in Model (12), as considered in CompLM (Wang et al., 2015; 2019a), the OLS-based estimates of ϖ and σ^2, denoted by $\widehat{\varpi}_{ols}$ and $\widehat{\sigma}^2_{ols}$, have the explicit solutions:

$$\widehat{\varpi}_{ols} = \left(\sum_{i=1}^{N} \mathbb{X}_i'\mathbb{X}_i\right)^{-1}\left(\sum_{i=1}^{N} \mathbb{X}_i'\boldsymbol{y}_i\right), \tag{14}$$

$$\widehat{\sigma}^2_{ols} = M^{-1}\sum_{i=1}^{N}(\boldsymbol{y}_i - \mathbb{X}_i\widehat{\varpi}_{ols})'(\boldsymbol{y}_i - \mathbb{X}_i\widehat{\varpi}_{ols}), \tag{15}$$

where $\mathbb{X}_i = (\boldsymbol{x}_i, \boldsymbol{u}_i\boldsymbol{W}_\mu, \boldsymbol{c}_i^*)$ for $i = 1, 2, \ldots, N$ and $M = \sum_{i=1}^{N} n_i$.

Under the normality assumption, the estimation procedure for the consistent numeric representations of complex data in Model (12) can be obtained within the multivariate framework of LMM. Laird et al. (1987) considered the full-sample likelihood function with respect to $\boldsymbol{y}$ and $\boldsymbol{\pi}_i$ $(i = 1, 2, \ldots, N)$ and derived the maximum likelihood (ML) estimates through the EM algorithm. Specifically, given a pair of estimates $\widehat{\boldsymbol{G}}^{(\omega)}$ and $(\widehat{\sigma}^{(\omega)})^2$, where the superscript ω indicates the iteration and $\omega = 0$ denotes the initial values, $\widehat{\varpi}^{(\omega)}$ is formulated as

$$\widehat{\varpi}^{(\omega)} = \left(\sum_{i=1}^{N} \mathbb{X}_i'\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)}\right)^{-1}\mathbb{X}_i\right)\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)}\right)\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)}\right)^{-1}\mathbb{X}_i\Bigg)^{-1}\left(\sum_{i=1}^{N} \mathbb{X}_i'\left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)}\right)^{-1}\boldsymbol{y}_i\right) \tag{16}$$

with

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)} = \mathbb{Z}_i \widehat{\boldsymbol{G}}^{(\omega)} \mathbb{Z}_i{}' + (\widehat{\sigma}^{(\omega)})^2 \boldsymbol{I}_{n_i} \tag{17}$$

and $\mathbb{Z}_i = (\boldsymbol{z}_i, \boldsymbol{v}_i \boldsymbol{W}_{q_\nu}, \boldsymbol{w}_i^*)$ for $i = 1, 2, \ldots, N$. On the contrary, when $\widehat{\varpi}^{(\omega)}$ is available, $\widehat{\boldsymbol{G}}^{(\omega)}$ and the others can be updated from $(\widehat{\sigma}^{(\omega)})^2$, that is,

$$\widehat{\boldsymbol{G}}^{(\omega+1)} = N^{-1} \sum_{i=1}^{N} \widehat{\boldsymbol{\pi}}_i^{(\omega)} \widehat{\boldsymbol{\pi}}_i^{(\omega)}{}' + \widehat{\boldsymbol{G}}^{(\omega)} - \widehat{\boldsymbol{G}}^{(\omega)} \mathbb{Z}_i{}' \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)} \right)^{-1} \mathbb{Z}_i \widehat{\boldsymbol{G}}^{(\omega)} \Bigg), \tag{18}$$

$$\left(\widehat{\sigma}^{(\omega+1)} \right)^2 = M^{-1} \sum_{i=1}^{N} \left((\widehat{\boldsymbol{e}}_i^{(\omega)})' \widehat{\boldsymbol{e}}_i^{(\omega)} + (\widehat{\sigma}^{(\omega)})^2 \mathrm{tr} \left(\boldsymbol{I}_{n_i} - (\widehat{\sigma}^{(\omega)})^2 \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)} \right)^{-1} \right) \right), \tag{19}$$

where

$$\widehat{\boldsymbol{\pi}}_i^{(\omega)} = \widehat{\boldsymbol{G}}^{(\omega)} \mathbb{Z}_i{}' \left(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)} \right)^{-1} (\boldsymbol{y}_i - \mathbb{X}_i \widehat{\varpi}^{(\omega)}), \tag{20}$$

$$\widehat{\boldsymbol{e}}_i^{(\omega)} = \boldsymbol{y}_i - \mathbb{X}_i \widehat{\varpi}^{(\omega)} - \mathbb{Z}_i \widehat{\boldsymbol{\pi}}_i^{(\omega)} \tag{21}$$

and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Such $\widehat{\boldsymbol{G}}^{(\omega+1)}$ and $(\widehat{\sigma}^{(\omega+1)})^2$ result in the update $\widehat{\varpi}^{(\omega+1)}$ from (16), which finishes an iteration.

Within the framework of the EM algorithm, the proposed procedure is always convergent due to the quadratic convex optimization involved. The convergence criterion is that the maximum difference between the present estimated parameters, say $\widehat{\varpi}^{(\omega)}$ and $(\widehat{\sigma}^{(\omega)})^2$, and the previous ones, say $\widehat{\varpi}^{(\omega-1)}$ and $(\widehat{\sigma}^{(\omega-1)})^2$, falls into a given threshold, say $\varepsilon = 0.01$, hat is,

$$\max \left\{ \left| \widehat{\varpi}^{(\omega)} - \widehat{\varpi}^{(\omega-1)} \right|_\infty, \left| (\widehat{\sigma}^{(\omega)})^2 - \left(\widehat{\sigma}^{(\omega-1)} \right)^2 \right| \right\} < \varepsilon, \tag{22}$$

where $\|\cdot\|_\infty$ denotes the maximum norm of a vector. Meanwhile, the algorithm also stops if it exceeds the iteration limit, say $l = 100$. As suggested by Laird et al. (1987), the initial value of $\widehat{\varpi}$, denoted by $\widehat{\varpi}^{(0)}$, is set to be $\widehat{\varpi}_{ols}$, and those of the other parameters can then be computed from $\widehat{\varpi}^{(0)}$ as

$$\widehat{\boldsymbol{G}}^{(0)} = N^{-1} \sum_{i=1}^{N} \left(\widehat{\boldsymbol{\pi}}_i^{(0)} \left(\widehat{\boldsymbol{\pi}}_i^{(0)} \right)' - \left(\widehat{\sigma}^{(0)} \right)^2 (\mathbb{Z}_i{}' \mathbb{Z}_i)^{-1} \right), \tag{23}$$

$$\left(\widehat{\sigma}^{(0)} \right)^2 = L^{-1} \sum_{i=1}^{N} \left(\boldsymbol{y}_i - \mathbb{Z}_i \widehat{\boldsymbol{\pi}}_i^{(0)} \right)' \left(\boldsymbol{y}_i - \mathbb{X}_i \widehat{\varpi}^{(0)} \right), \tag{24}$$

where $L = M - (N-1)q - p$ and $\widehat{\boldsymbol{\pi}}^{(0)} = (\mathbb{Z}_i{}' \mathbb{Z}_i)^{-1} \mathbb{Z}_i{}' (\boldsymbol{y}_i - \mathbb{X}_i \widehat{\varpi}^{(0)})$. The aforementioned initialization for the estimation procedure begins with the reduced OLS-based linear regression, and further abstracts the subject-specific information from the covariance structure of errors.

Remark 4. The proposed parameter estimation for CompLMM through the EM algorithm is consistent with the existing solutions for CompLM in (14) and (15) proposed by Wang et al. (2015, 2019a). Actually, when there are no random effect, namely no z_{ijk}, ν_{ijk} and $\boldsymbol{w}_{ijk}$ in Model (1), $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)}$ reduces to be proportional to $\boldsymbol{I}_{n_i}$ with no $\mathbb{Z}_i$ involved in (17), implying that $\widehat{\varpi}^{(\omega)}$ in (16) identically equals $\widehat{\varpi}_{ols}$ in (14) for any ω. A similar conclusion can also be drawn on (19), namely $(\widehat{\sigma}^{(\omega+1)})^2 \equiv \widehat{\sigma}_{ols}^2$, since $\widehat{\boldsymbol{e}}_i^{(\omega)} = \boldsymbol{y}_i - \mathbb{X}_i \widehat{\varpi}^{(\omega)}$ and $(\widehat{\sigma}^{(\omega)})^2 (\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)})^{-1} = \boldsymbol{I}_{n_i}$. We conclude from these results that CompLM works exactly as the pooled method for longitudinal complex data.

As a result of the ML estimation, the aforementioned estimates for these transformed data enjoy the properties of consistence and asymptotic normality. Since the truncated expansion coefficients and ilr coordinates represent the original compositional and functional data equivalently or approximately equivalently, the properties for their ML estimates may also indicate those for the original complex data with diversified characteristics. Actually, as discussed in Remarks 1–3, the distributions on multiple types of complex data are defined via their representations; therefore, the ML estimation procedure on the represented data in Model (12) can be exactly regarded as on the original complex data with diversified characteristics in Model (1). In summary, we present the computational procedure of the proposed method for longitudinal complex data in Algorithm 1.

Algorithm 1. Computational procedure for CompLMM

Input: The data set $\left\{(y_{ij}, x_{ijk}, z_{ijk}, o^m_{\mu_{ijk}}, o^m_{\nu_{ijk}}, \boldsymbol{c}_{ijk}, \boldsymbol{w}_{ijk}; t^m_{\mu_{ijk}}, t^m_{\nu_{ijk}})\right\}^{N,n_i,p_*,n}_{i,j,k,m=1}$, with p_* corresponding to the related dimension, including the responses $\{y_{ij}\}^{N,n_i}_{i,j=1}$, scalar covariates $\left\{(x_{ijk}, z_{ijk})\right\}^{N,n_i,p_x/q_z}_{i,j,k=1}$, observations from functional covariates $\{(o^m_{\mu_{ijk}}, o^m_{\nu_{ijk}})\}^{N,n_i,p_\mu/p_\nu,n}_{i,j,k,m=1}$ at times $\{(t^m_{\mu_{ijk}}, t^m_{\nu_{ijk}})\}^{N,n_i,p_\mu/q_\nu,n}_{i,j,k,m=1}$ and compositional covariates $\{(\boldsymbol{c}_{ijk}, \boldsymbol{w}_{ijk})\}^{N,n_i,p_c/q_w}_{i,j,k=1}$; the given K basis functions $\{\phi_i\}^K_{i=1}$; the initial value of the parameter $\widehat{\boldsymbol{\varpi}}^{(0)}$, associated with the intermediate $\widehat{\boldsymbol{\Sigma}}^{(0)}_{\boldsymbol{y}_i}$; the convergence threshold ϵ; and the iteration limit l.

Output: $\widehat{\boldsymbol{\Theta}} = \{\widehat{\boldsymbol{\varpi}}, \widehat{\boldsymbol{G}}, \hat{\sigma}^2\}$ and $\widehat{\boldsymbol{\pi}}_i$ for $i = 1, 2, ..., N$.

1: Compute the expansion coefficients $\boldsymbol{u}_{ijk}$ and $\boldsymbol{v}_{ijk}$ $(i = 1, 2, ..., N; j = 1, 2, ..., n_i)$:

$$\boldsymbol{u}_{ijk} = (\boldsymbol{\Phi}'_{\mu_{ijk}}\boldsymbol{\Phi}_{\mu_{ijk}})^{-1}\boldsymbol{\Phi}'_{\mu_{ijk}}\boldsymbol{o}_{\mu_{ijk}} \quad (k = 1, 2, ..., p_\mu),$$
$$\boldsymbol{v}_{ijk} = (\boldsymbol{\Phi}'_{\nu_{ijk}}\boldsymbol{\Phi}_{\nu_{ijk}})^{-1}\boldsymbol{\Phi}'_{\nu_{ijk}}\boldsymbol{o}_{\nu_{ijk}} \quad (k = 1, 2, ..., p_\nu),$$

where the notations coincide with (2) and the subscript indicates the functional covariate;

2: compute the ilr coordinates $\boldsymbol{c}^*_{ijk}$ and $\boldsymbol{w}^*_{ijk}$ from (8);

3: construct the data matrices $\mathbb{X}_i$ and $\mathbb{Z}_i$ $(i = 1, 2, ..., N)$; set $\omega = 0$;

4: **repeat**

5: compute the intermediates $\widehat{\boldsymbol{\pi}}^{(\omega+1)}_i$ and $\widehat{\boldsymbol{e}}^{(\omega+1)}_i$ $(i = 1, 2, ..., N)$ from (20) and (21), respectively;

6: update $\widehat{\boldsymbol{G}}^{(\omega+1)}$ and $(\hat{\sigma}^{(\omega+1)})^2$ from (18) and (19), respectively;

7: update the intermediate $\widehat{\boldsymbol{\Sigma}}^{(\omega+1)}_{\boldsymbol{y}_i}$ $(i = 1, 2, ..., N)$ from (17);

8: update $\widehat{\boldsymbol{\varpi}}^{(\omega+1)}$ from (16);

9: let $\omega := \omega + 1$;

10: **until** (22) holds or $\omega > l$;

11: **return**

$$\widehat{\boldsymbol{\Theta}} := \{\widehat{\boldsymbol{\varpi}}^{(t)}, \widehat{\boldsymbol{G}}^{(t)}, (\hat{\sigma}^{(t)})^2\} \quad \text{and} \quad \widehat{\boldsymbol{\pi}}_i := \widehat{\boldsymbol{\pi}}^{(t)}_i \ (i = 1, 2, ..., N).$$

3.3. Some issues

In this subsection, we discuss some other issues in both theory and implementation for the proposed CompLMM.

First, when the basis functions used are not orthonormal, such as the B-spline functions, the related metric matrix $\boldsymbol{W}$ will not be the identical matrix. To compute it numerically, one may first uniformly take T samples, say $\{\tau_1, \tau_2, \ldots, \tau_T\}$, from $\mathcal{I}$, and then approximate it as

$$\boldsymbol{W} = T^{-1}\sum_{i=1}^{T}\phi(\tau_i)\phi'(\tau_i).$$

By contrast, one may also generate a new group of orthonormal basis functions from those previously given such that the related metric matrix will naturally be identical. In this study, we perform the functional PCA (FPCA) procedure, as it is a popular method for conducting the orthonormal basis function system in FDA (Ramsay & Silverman, 2005).

Specifically, suppose that we have M groups of expansion coefficients $\boldsymbol{u}_{ij} = (u_{ij1}, u_{ij2}, \ldots, u_{ij,n_i})'$ for μ_{ij} $(i = 1, 2, \ldots, N; j = 1, 2, \ldots, n_i)$ under the given basis function system ϕ, we wish to generate K orthonormal basis functions from ϕ, say $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K)'$. Let $\boldsymbol{U} = (\boldsymbol{u}_{11}, \boldsymbol{u}_{12}, \ldots, \boldsymbol{u}_{21}, \ldots, \boldsymbol{u}_{N,n_N})'$ and $\boldsymbol{L}$ denote the Cholesky decomposition of $\boldsymbol{W}$ such that $\boldsymbol{W} = \boldsymbol{L}\boldsymbol{L}'$, we consider the SVD problem $M^{-1}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{W}\boldsymbol{b} = \lambda\boldsymbol{b}$ subject to $\boldsymbol{b}'\boldsymbol{W}\boldsymbol{b} = 1$, or equivalently the PCA problem with $\boldsymbol{b}^* = \boldsymbol{L}'\boldsymbol{b}$:

$$M^{-1}\boldsymbol{L}'\boldsymbol{U}'\boldsymbol{U}\boldsymbol{L}\boldsymbol{b}^* = \lambda \boldsymbol{b}^* \tag{25}$$

Subject to $(\boldsymbol{b}^*)'\boldsymbol{b}^* = 1$. Therefore, the orthonormal basis functions and related expansion coefficients for μ_{ij}, denoted as $\boldsymbol{u}_{ij}^*$, can be expressed as

$$\xi_k = \boldsymbol{b}_k{}'\phi \quad (k=1,2,\ldots,K) \quad \text{and} \quad \boldsymbol{u}_{ij}^* = \boldsymbol{B}'\boldsymbol{W}\boldsymbol{u}_{ij},$$

Respectively, where $\boldsymbol{b}_k$ corresponds to the eigenvector with the k-th largest eigenvalue from (25) and $\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_K)$. For more theoretical details of FPCA, see Ramsay and Silverman (2005).

In practice, given no information on the proper basis function system, we shall first expand the observations under a relatively large number, say $K_0 > K$, of basis functions, and then perform the FPCA procedure to summarize them into K orthonormal basis functions. Meanwhile, through the PCA-based estimation in FPCA, we shall introduce the cumulative contribution rate (CCR) of variance to determine the dimension of the orthonormal basis function system, namely K. Let λ_k for $k = 1, 2, \ldots, K_0$ be the descending-ordered eigenvalues from (25), we set K to be the minimum value that reaches the threshold of CCR denoted by δ, namely

$$K = \min\left\{ K : \sum_{k=1}^{K} \lambda_k \Big/ \sum_{k=1}^{K_0} \lambda_k \geq \delta \right\}.$$

Empirically, $\delta \geq 0.85$ would lead to a reasonable orthonormal basis function system that preserves most of the informative variation of the observations. As indicated by the simulation results, the orthonormal basis functions along with the related expansion coefficients through the FPCA procedure efficiently support the proposed method.

Then, we consider the statistical inferences on the estimated coefficients of the fixed effects with diversified characteristics. Under the normality assumption on the error, the covariance matrix of $\widehat{\varpi}$ in Model (12) can be derived as

$$\mathrm{Cov}(\widehat{\varpi}) = \left(\sum_{i=1}^{N} \mathbb{X}_i{}' \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{-1} \mathbb{X}_i \right)^{-1}, \tag{26}$$

where the diagonal elements correspond to the estimated variances of the consistent numeric representations of complex data. From (26), the statistical inferences on the scalar covariates can be directly built up. Specifically, for $\widehat{\alpha}_k$ and its variance $\widehat{\sigma}^2(\alpha_k)$ $(k = 1, 2, \ldots, p_x)$, $\widehat{\alpha_k}/\widehat{\sigma}(\alpha_k)$ would asymptotically satisfy the standard normal distribution when the null hypothesis holds, namely $H_0 : \alpha_k = 0$. Similar conclusions can be drawn on the estimated results for functional and compositional covariates, however, the significance of the estimated expansion coefficients and ilr coordinates are meaningless, since we focus on the functional and compositional structures instead of their numeric representations. Thus, we shall develop the statistical inferences on the original function in a point-wise manner and the clr-transformed composition.

For the estimated functional coefficient $\widehat{\beta}_k$ $(k = 1, 2, \ldots, p_\mu)$, the related estimate of the variance function can be expressed as

$$\widehat{\mathcal{K}}_{\beta_k}(t) = \boldsymbol{\xi}'(t)\mathrm{Cov}(\widehat{\boldsymbol{\lambda}}_k)\boldsymbol{\xi}(t) = \phi'(t)\boldsymbol{B}\mathrm{Cov}(\widehat{\boldsymbol{\lambda}}_k)\boldsymbol{B}'\phi(t) \qquad (t \in \mathscr{T}),$$

where $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K)'$ comprises the orthonormal basis functions generated from ϕ, and $\widehat{\boldsymbol{\lambda}}_k$ is the expansion coefficients of $\widehat{\beta}_k$ under $\boldsymbol{\xi}$. Therefore, we shall obtain the confidence band of $\widehat{\beta}_k$ in a point-wise manner as

$$\left(\widehat{\beta}_k(t) + Z_{(\varrho/2)}\sqrt{\widehat{\mathcal{K}}_{\beta_k}(t)},\ \widehat{\beta}_k(t) + Z_{(1-\varrho/2)}\sqrt{\widehat{\mathscr{K}}_{\beta_k}(t)} \right) \qquad (t \in \mathscr{T}),$$

where $Z_{(\varrho/2)}$ and $Z_{(1-\varrho/2)}$ denote the $\varrho/2$ and $1-\varrho/2$ quantiles of the standard normal distribution, respectively, for a given significance level ϱ. For the estimated compositional coefficient $\widehat{\boldsymbol{\gamma}}_k$, the direct statistical inferences on the compositional structure would be difficult due to the uncertain magnitude of scale in the closure operation. An alternative solution is to derive the statistical inferences on the clr-transformed composition as they also carry the relative information of the parts in the original composition. Since $\mathrm{clr}(\widehat{\boldsymbol{\gamma}}_k) = \boldsymbol{\Psi}'\mathrm{ilr}(\widehat{\boldsymbol{\gamma}}_k)$ $(k = 1, 2, \ldots, p_c)$, the estimate of the covariance matrix is expressed as $\mathrm{Cov}(\mathrm{clr}(\widehat{\boldsymbol{\gamma}}_k)) = \boldsymbol{\Psi}'\mathrm{Cov}(\mathrm{ilr}(\widehat{\boldsymbol{\gamma}}_k))\boldsymbol{\Psi}$. Thus, for $\mathrm{clr}_i(\widehat{\boldsymbol{\gamma}}_k)$ $(i = 1, 2, \ldots, D)$ and related variance $\mathrm{Var}(\mathrm{clr}_i(\widehat{\boldsymbol{\gamma}}_k))$, we shall compare it with 0 through $\mathrm{clr}_i(\widehat{\boldsymbol{\gamma}}_k)/\sqrt{\mathrm{Var}(\mathrm{clr}_i(\widehat{\boldsymbol{\gamma}}_k))}$ to evaluate whether the i-th part in the composition is significantly above/below the average level among all the parts under the significance level ϱ. Similarly, the statistical inferences for the reduced CompLM can be derive from $\mathrm{Cov}(\widehat{\varpi}) = (\sum_{i=1}^{N} \mathbb{X}_i{}'\mathbb{X}_i)^{-1}$.

Next, we exemplify the proposed framework for longitudinal complex data using three types of covariates, and this framework is available for more types of complex data with diversified characteristics. For example, introducing dummy variables is common for processing categorical, nominal and ordinal variables in longitudinal analysis (Hsiao, 2014). We can represent these as groups of dummy variables and conduct compLMM for these multiple scalar covariates. For unstructured

text data, we can summary these into a series of compositions associated with the frequencies of topics/positions, similar to the measure of investors' sentiments by Zhou et al. (2018), and therefore analyze text data under the proposed framework.

The key technique for formulating CompLMM is to find the suitable representation for a specific type of complex data and suitable algebraic system, such as the basis function expansion with the l_2 norm for the Hilbert function space in FDA, and ilr transformation with the Aitchison inner product for the simplex in CDA. Following this idea, more diversified types of variables could be jointed into the proposed framework. For example, in the context of symbolic data analysis (SDA), the interval-valued variable has special binary representations such as "Lower-Upper" and "Center-Radius" (Billard & Diday, 2003; Sun et al., 2018). Linear regression can then be conducted on these binary numeric variables (Wei et al., 2017), with the random effects incorporated analogously. Similarly, we can also formulate the regressions on the other symbolic variables in SDA, namely histograms and distribution functions, with more complicated characteristics based on the Wasserstein distance (Irpino & Verde, 2015), and consider related random effects to extend them to the proposed framework.

Finally, the introduction of random effects promotes the performance of linear regression for longitudinal complex data, while the complexity of random effects may also lead to an extra cost of computation and a loss of degrees of freedom. Thus, the trade-off between the improvement in fitting accuracy and complexity of random effects is worthy of consideration, which falls into the selection of random effects. As an important issue for LMM, many statistical solutions for traditional scalar covariates have been proposed, such as the Bayesian information criteria selector (Fitzmaurice et al., 2011) and joint selection (Bondell et al., 2010). Furthermore, we can determine the constitution of the random effects from the practical and empirical perspectives (e.g., some financial knowledge in the real data study). We can also conduct a series of alternative CompLMM associated with all the possible constitutions of the random effects, including CompLM, and select the balanced one that approximates the best improvement with relatively few random effects.

4. Simulation studies

In this section, we report the simulation results to evaluate the performance of the proposed parameter estimation for CompLMM. Three measures are introduced: the squared ratio error (SRE) for scalar responses, integral squared error (ISE) for functions, and absolute ratio error (ARE) for compositions. These are respectively defined in Model (1) as

$$\mathrm{SRE} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left(y_{ij}-\widehat{y}_{ij}\right)^2 \Big/ \sum_{i=1}^{N}\sum_{j=1}^{n_i} y_{ij}^2,$$

$$\mathrm{ISE}(\widehat{\beta}_k) = d^2_{\mathcal{L}^2}(\beta_k, \widehat{\beta}_k) \quad (k=1,2,\ldots,p_\mu),$$

$$\mathrm{ARE}(\widehat{\gamma}_k) = d_a(\gamma_k, \widehat{\gamma}_k)/\parallel \gamma_k\parallel_a \quad (k=1,2,\ldots,p_c),$$

where $\widehat{y}_{ij}$, $\widehat{\mu}_{ijk}$ and $\widehat{\boldsymbol{c}}_{ijk}$ denote the related fitted values. A lower value of SRE, ISE or ARE indicates a more accurate fitting for the specific response, function or composition, respectively.

We generate the data from the following model:

$$y_{ij} = 2 + \alpha_1 x_{ij1} + \int_0^1 \beta_1\mu_{ij1} + \int_0^1 \beta_2\mu_{ij2} + \left(\boldsymbol{\gamma}_1, \boldsymbol{c}_{ij1}\right)_a + \left(\boldsymbol{\gamma}_2, \boldsymbol{c}_{ij2}\right)_a + a_{i0} + \int_0^1 b_{i1}\nu_{ij1} + \left(\boldsymbol{r}_{i1}, \boldsymbol{w}_{ij1}\right)_a + \varepsilon_{ij},$$

where seven cubic B-spline basis functions defined by four equally spaced inner knots over $[0,1]$, say $\phi = \{\phi_1,\phi_2,\ldots,\phi_7\}$, and the ilr coordinates from (8) and (9) are adopted.

The detailed settings of the other parameters are introduced as follows.

a) In the fixed effects, x_{ij1} is independently generated from the standard normal distribution, with $\alpha_1 = 5$; β_1 and β_2 are linearly combined by ϕ, with the symmetric combination coefficients. That is, for any $t\in[0,1]$,

$$\beta_1(t) = \sum_{j=1}^{7}(4-j)\phi_j(t) \quad \text{and} \quad \beta_2(t) = \sum_{j=1}^{7}(j-4)\phi_j(t);$$

Respectively; μ_{ij1} and μ_{ij2} are described as $n = 200$ samples observed at times $\{t_1, t_2, \ldots, t_n\}$ from the linear combinations of ϕ with measurement errors, that is,

$$\mu_{ijk}(t_l) = \boldsymbol{u}_{ijk}{}'\phi(t_l) + \varepsilon_{ijkl} \quad (k=1,2;\ l=1,2,\ldots,n),$$

where the expansion coefficients of two functions are sampled from $\mathcal{N}(0_7, \boldsymbol{I}_7)$, and the errors are generated from $\mathcal{N}(0, 0.1^2)$; $\boldsymbol{c}_{ij1}$ and $\boldsymbol{c}_{ij2}$ are separately generated from the simplicial normal distribution $\mathcal{N}_S(0_2, \boldsymbol{I}_2)$ (Mateu-Figueras et al., 2013), with the compositional coefficients $\boldsymbol{\gamma}_1 = [0.6, 0.2, 0.2]'$ and $\boldsymbol{\gamma}_2 = [0.25, 0.25, 0.5]'$ in S^3, respectively.

b) In the random effects, the covariates are constituted by the intercept a_{i0} and the first function and composition, namely $\nu_{ij1} = \mu_{ij1}$ and $\boldsymbol{w}_{ij1} = \boldsymbol{c}_{ij1}$; b_{i1} and $\boldsymbol{r}_{i1}$ are represented by the expansion coefficients under ϕ and ilr coordinates, say $\boldsymbol{\theta}_{i1}$ and $\boldsymbol{r}^*_{i1}$, respectively. The parameters, namely $\boldsymbol{\pi}_i = (a_{i0}; \boldsymbol{\theta}_{i1}; \boldsymbol{r}^*_{i1})$, are then jointly generated from $\mathcal{N}(0_{10}, \boldsymbol{G})$, where $\boldsymbol{G}$ is blocked diagonal as $\boldsymbol{G} = \mathrm{diag}(9, \boldsymbol{G}_{\theta^*}, 0.5\boldsymbol{I}_4, \boldsymbol{G}_r)$ with

$$\boldsymbol{G}_{\theta^*} = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{G}_r = \begin{pmatrix} 9 & 4.8 \\ 4.8 & 4 \end{pmatrix}.$$

c) ε_{ij} is independently generated from $\mathcal{N}(0, \sigma^2)$, where σ takes a value of 0.5, 1 and 1.5 to reflect the signal-to-noise ratio (SNR) from strong to weak.

Three combinations of the number of individuals N and sample size for each individual n_i are considered: $(N, n_i) = (100, 30)$, $(100, 60)$ and $(300, 60)$. For each case, we independently replicate the simulation 500 times and conduct the proposed CompLMM as well as the baseline CompLM for comparison. Specifically, to mimic the lack of information in practice on the oracle basis function system, we perform the FPCA procedure to determine the basis functions. Specifically, the previously given basis function system is set to consist of the cubic B-spline functions generated by 17 equally spaced inner knots over $[0, 1]$, namely $K_0 = 20$. The numbers of the orthonormal basis functions for two functional covariates through the FPCA procedure are determined by the CCR-based criteria with $\delta = 0.9$. Table 1 reports the estimated results for the fixed effects of scalar and compositional covariates, Table 2 summarizes the performance of the regression model, and Fig. 1 then visualizes the curves of the estimated coefficients of the fixed effects for two functional covariates in some replications selected randomly.

As shown in Table 1, the estimated coefficients of the fixed effects for the scalar covariates obtained from both CompLMM and CompLM approximate the related ideal values on average, while those from CompLMM are more stable with lower

Table 1
Means and standard derivations (in brackets) of the estimated coefficients for the fixed effects of scalar and compositional covariates. The row "True" denotes the related true values.

(N, n_i)	Model	Scalar		Compositional					
		Intercept	$\widehat{\alpha}_1$	$\widehat{c}_{11}$	$\widehat{c}_{12}$	$\widehat{c}_{13}$	$\widehat{c}_{21}$	$\widehat{c}_{22}$	$\widehat{c}_{23}$
$\sigma = 0.5$									
(100, 30)	CompLMM	1.986 (0.379)	5 (0.017)	0.598 (0.09)	0.197 (0.027)	0.205 (0.071)	0.25 (0.002)	0.25 (0.002)	0.5 (0.003)
	CompLM	1.982 (0.381)	5.004 (0.173)	0.596 (0.095)	0.197 (0.033)	0.206 (0.074)	0.25 (0.021)	0.251 (0.023)	0.499 (0.029)
(100, 60)	CompLMM	1.99 (0.397)	5 (0.012)	0.6 (0.085)	0.197 (0.026)	0.202 (0.067)	0.25 (0.002)	0.25 (0.002)	0.5 (0.002)
	CompLM	1.987 (0.401)	5.001 (0.117)	0.599 (0.087)	0.197 (0.029)	0.203 (0.069)	0.249 (0.016)	0.251 (0.017)	0.5 (0.021)
(300, 60)	CompLMM	1.978 (0.242)	5 (0.007)	0.597 (0.051)	0.2 (0.016)	0.203 (0.04)	0.25 (0.001)	0.25 (0.001)	0.5 (0.001)
	CompLM	1.979 (0.242)	4.993 (0.069)	0.598 (0.053)	0.2 (0.019)	0.203 (0.04)	0.25 (0.01)	0.25 (0.009)	0.499 (0.012)
$\sigma = 1$									
(100, 30)	CompLMM	1.986 (0.378)	5 (0.034)	0.598 (0.09)	0.197 (0.027)	0.205 (0.072)	0.25 (0.005)	0.25 (0.005)	0.5 (0.006)
	CompLM	1.982 (0.38)	5.003 (0.175)	0.596 (0.095)	0.198 (0.033)	0.206 (0.075)	0.25 (0.021)	0.25 (0.023)	0.5 (0.029)
(100, 60)	CompLMM	1.99 (0.397)	5 (0.024)	0.601 (0.085)	0.197 (0.027)	0.202 (0.067)	0.25 (0.003)	0.25 (0.003)	0.5 (0.004)
	CompLM	1.987 (0.402)	5.001 (0.12)	0.599 (0.087)	0.197 (0.029)	0.203 (.069)	0.249 (0.016)	0.251 (0.017)	0.5 (0.021)
(300, 60)	CompLMM	1.978 (0.242)	5 (0.014)	0.597 (0.052)	0.2 (0.016)	0.203 (0.04)	0.25 (0.002)	0.25 (0.002)	0.5 (0.002)
	CompLM	1.979 (0.242)	4.993 (0.07)	0.598 (0.053)	0.2 (0.02)	0.203 (0.004)	0.251 (0.01)	0.25 (0.009)	0.499 (0.012)
$\sigma = 1.5$									
(100, 60)	CompLMM	1.987 (0.378)	5 (0.051)	0.597 (0.091)	0.197 (0.027)	0.205 (0.072)	0.25 (0.007)	0.25 (0.007)	0.5 (0.009)
	CompLM	1.982 (0.38)	5.003 (0.178)	0.596 (0.095)	0.198 (0.033)	0.206 (0.075)	0.25 (0.022)	0.25 (0.024)	0.5 (0.03)
(100, 60)	CompLMM	1.99 (0.398)	5 (0.036)	0.601 (0.085)	0.197 (0.027)	0.202 (0.067)	0.25 (0.005)	0.25 (0.005)	0.5 (0.006)
	CompLM	1.987 (0.402)	5.001 (0.123)	0.599 (.087)	0.198 (0.029)	0.203 (0.069)	0.249 (0.016)	0.251 (0.017)	0.5 (0.022)
(300, 60)	CompLMM	1.978 (0.242)	5 (0.02)	0.597 (0.052)	0.2 (0.017)	0.203 (0.04)	0.25 (0.003)	0.25 (0.003)	0.5 (0.004)
	CompLM	1.979 (0.242)	4.993 (0.072)	0.598 (0.053)	0.2 (0.02)	0.203 (0.04)	0.251 (0.01)	0.25 (0.01)	0.499 (0.012)
True		2	5	0.6	0.2	0.2	0.25	0.25	0.5

Table 2
Means and standard derivations (in brackets) of $\widehat{\sigma}^2$ and three measures SRE, ISE and ARE.

(N, n_i)	Model	Functional		Compositional		$\widehat{\sigma}^2$	SRE
		ISE($\widehat{\boldsymbol{\beta}}_1$)	ISE($\widehat{\boldsymbol{\beta}}_2$)	ARE($\widehat{\boldsymbol{c}}_1$)	ARE($\widehat{\boldsymbol{c}}_2$)		
$\sigma = 0.5$							
(100, 30)	CompLMM	0.043 (0.019)	0.029 (0.008)	0.352 (0.219)	0.022 (0.012)	0.253 (0.009)	0.17 (0.013)
	CompLM	0.882 (0.667)	0.909 (0.668)	0.38 (0.23)	0.209 (0.112)	29.031 (2.872)	0.67 (0.035)
(100, 60)	CompLMM	0.037 (0.016)	0.024 (0.004)	0.336 (0.222)	0.015 (0.008)	0.251 (0.005)	0.02 (0.015)
	CompLM	0.458 (0.327)	0.474 (0.368)	0.353 (0.218)	0.152 (0.083)	29.204 (2.767)	0.672 (0.034)
(300, 60)	CompLMM	0.037 (0.014)	0.033 (0.012)	0.201 (0.128)	0.025 (0.013)	2.247 (0.024)	0.05 (0.005)
	CompLM	0.174 (0.112)	0.175 (0.119)	0.207 (0.126)	0.089 (0.046)	29.339 (1.629)	0.678 (0.021)
$\sigma = 1$							
(100, 30)	CompLMM	0.07 (0.036)	0.056 (0.026)	0.355 (0.22)	0.043 (0.023)	1.005 (0.054)	0.032 (0.013)
	CompLM	0.9 (0.671)	0.933 (.675)	0.382 (0.231)	0.212 (0.112)	29.783 (2.877)	0.676 (0.034)
(100, 60)	CompLMM	0.049 (0.021)	0.036 (0.013)	0.337 (0.222)	0.03 (0.016)	0.999 (0.02)	0.035 (0.014)
	CompLM	0.467 (0.339)	0.49 (0.379)	0.353 (0.218)	0.154 (0.084)	29.954 (2.77)	0.678 (0.033)
(300, 60)	CompLMM	0.03 (0.011)	0.026 (0.009)	0.2 (0.128)	0.017 (0.009)	0.999 (0.011)	0.025 (0.004)
	CompLM	0.179 (0.116)	0.179 (0.119)	0.208 (0.126)	0.09 (0.047)	30.089 (1.627)	0.684 (0.021)
$\sigma = 1.5$							
(100, 30)	CompLMM	0.115 (0.073)	0.098 (.057)	0.358 (0.221)	0.065 (0.034)	2.275 (0.163)	0.057 (0.014)
	CompLM	0.931 (0.687)	0.971 (0.694)	0.385 (0.232)	0.217 (0.114)	31.032 (2.885)	0.685 (0.033)
(100, 60)	CompLMM	0.068 (0.034)	0.056 (0.028)	0.337 (0.222)	0.045 (0.023)	2.246 (0.044)	0.06 (0.014)
	CompLM	0.483 (0.356)	0.514 (0.396)	0.354 (0.218)	0.157 (0.085)	31.203 (2.773)	0.687 (0.032)
(300, 60)	CompLMM	0.037 (0.014)	0.033 (0.012)	0.201 (0.128)	0.025 (.013)	2.247 (0.024)	0.05 (0.005)
	CompLM	0.186 (0.122)	0.185 (0.121)	0.208 (0.126)	0.092 (0.048)	31.338 (1.627)	0.693 (0.02)

standard deviations than the baseline, for example, 0.017 (CompLMM) vs. 0.173 (CompLM) for $\widehat{\alpha}_1$ with $(N, n_i) = (100, 30)$ and $\sigma = 0.5$. As the sample size increased, the estimated results got stable, with decreasing standard deviations.

Then, for the functional covariates, as shown in Table 2, CompLMM sharply improves the estimation efficiency of two overall functions, where the ISE values for two functional covariates extremely approach zero, with generally small standard deviations. The result also indicates that the orthonormal basis function system and related expansion coefficients through the FPCA procedure are reliable and the FPCA procedure can efficiently support the proposed CompLMM. By contrast, CompLM performs relatively bad in estimating the functional covariates here. When the sample size is small, say $(N, n_i) = (100, 30)$, the ISE values for CompLM approach almost to 1, with large standard derivations. When the sample size gets larger, say $(N, n_i) = (300, 60)$, the ISE values decrease gradually to 0.1 above, which is still multiple times larger than those for CompLMM.

Next, for the compositional covariates, as reported in Tables 1 and 2, the two methods perform almost the same, where the estimates for most components of two compositional covariates and the ARE values for CompLMM are slightly more stable than those for ComLM.

Finally, for the fitting performance of the regression summarized in Table 2, CompLMM well estimates σ for different levels of SNR, whereas the estimates of CompLM far exceed the corresponding true values. Moreover, CompLM almost fails to fit the responses, since the related SRE values are on average above 0.6; by contrast, for CompLMM, the average values of SRE are significantly low and close to 0.

As exemplified in Fig. 1, the curves (in red) of the estimated functional coefficients of the fixed effects by CompLMM get closer to the true settings (in grey) than those (in cyan) by CompLM. For both the increasing and decreasing cases for the functional coefficients, CompLMM fits the functions well across the interval, whereas CompLM, although capturing the general trends of the functions, involves relatively large periodic perturbations when the sample size is small, for example, $(N, n_i) = (100, 30)$. Moreover, the estimated biases between the true and fitted curves from the two models are eliminated gradually as the sample size increases to $(N, n_i) = (300, 60)$. Specifically, the estimated biases for the first function having the random effects are relatively larger than those for the second one having no random effects.

In summary, the proposed CompLMM succeeds in addressing the longitudinal features within complex data with diversified characteristics, especially those with functional characteristics.

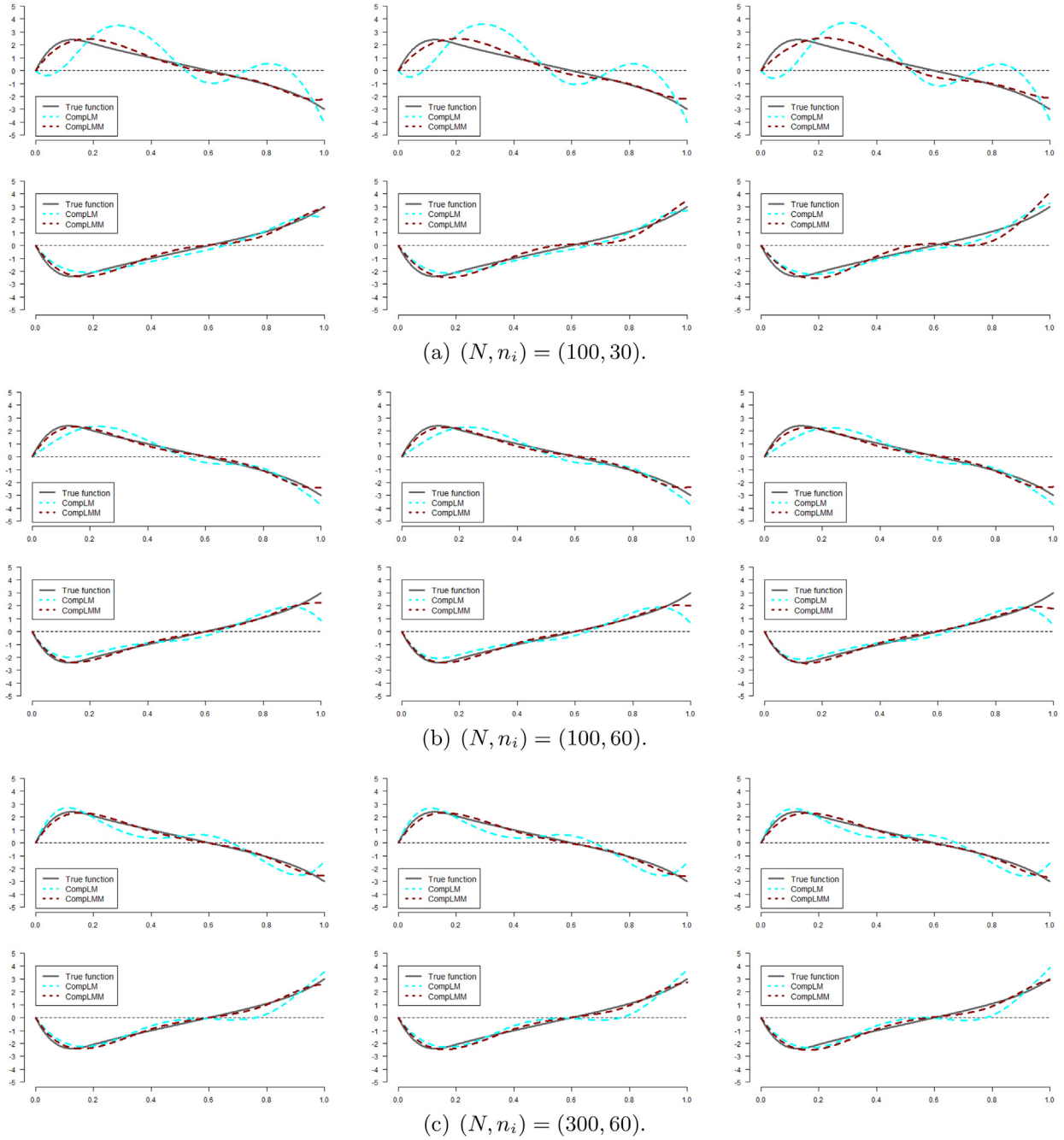


Fig. 1. Curves of the estimated functional coefficients. The column panels from left to right denote the three levels of SNR from $\sigma = 0.5$ to $\sigma = 1.5$. The upper and lower sub-rows indicate the first and second functional covariates, respectively.

5. Application to Chinese stock market

In this section, we adopt the proposed CompLMM in the example of Chinese stock market to demonstrate its usefulness. The existing approach to complex data modeling, namely CompLM (Wang et al., 2015; 2019a), is also used for comparison. Here, we aim at the influence of the indirect information as well as the historical price trend on the stocks price. As exemplified by numerous studies, the macroeconomic indicators (Chen et al., 1986), public online sentiment (Ruan et al., 2018; Zhou et al., 2018) and analysts' recommendations (Duan et al., 2013) may improve the interpretability and accuracy of the models for this problem.

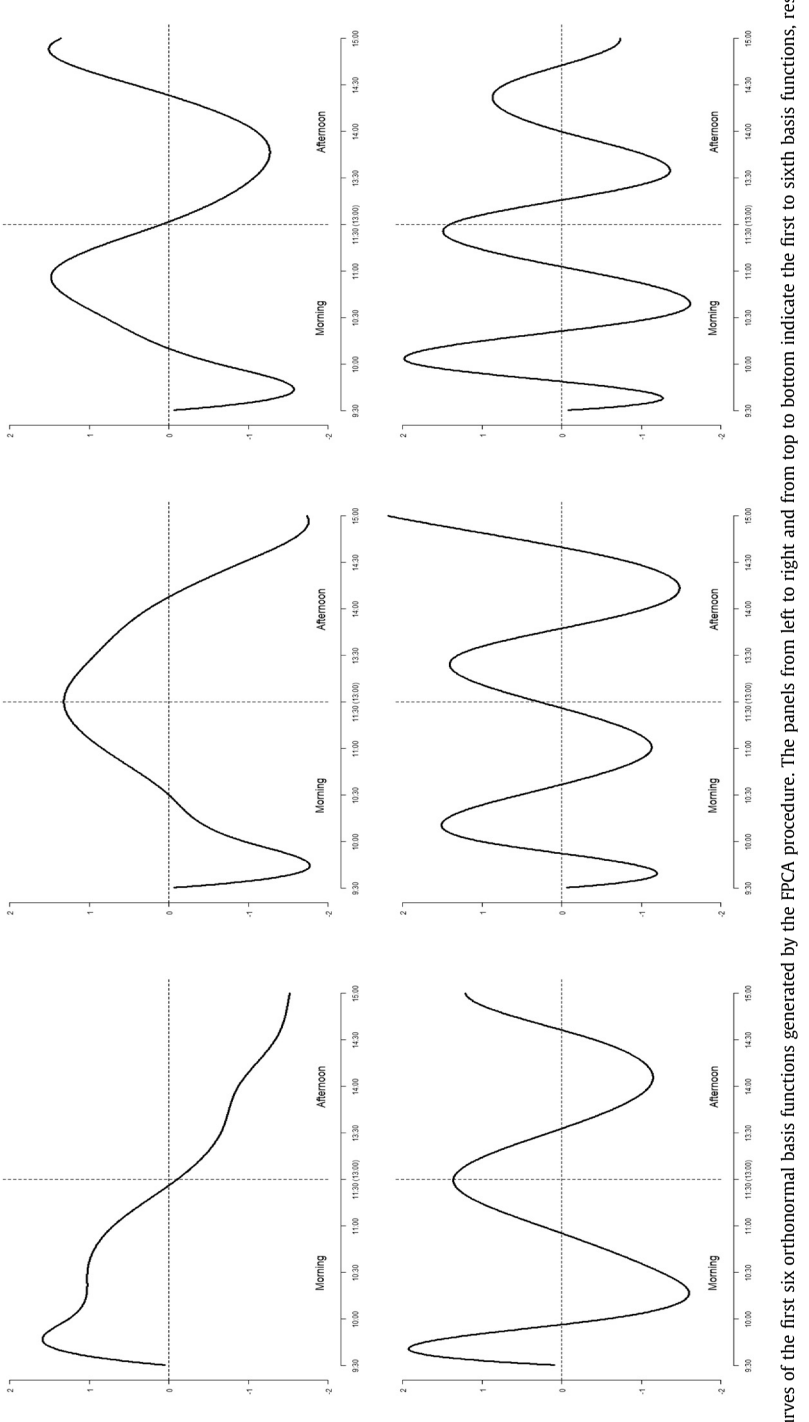


Fig. 2. Curves of the first six orthonormal basis functions generated by the FPCA procedure. The panels from left to right and from top to bottom indicate the first to sixth basis functions, respectively.

In this case, we regress the daily closing price (DCP) of stocks against the related daily volume (DV), intraday percentage return (IPR) and online investors' sentiments (OIS) in the former session. Data on the constituent stocks in CSI300 from January 8 to April 29, 2016 (75 trading days) are collected from the Wind service. The stocks with prices over 50 or active trading days fewer than 40 are removed, and finally 256 stocks are left. Here, both DCP and DV are scalar, IPR recorded every 5 min is described as a series of smoothing curves from opening to closing, and OIS is measured naturally as a compositional structure associated with five types of sentiments labeled "Anger", "Disgust", "Joy", "Sadness" and "Fear" (Zhou et al., 2018). Thus, the regression model contains two scalar covariates (including the intercept), one functional covariate and one compositional covariate.

In the FPCA procedure, the previously given basis system consists of ten cubic B-spline functions determined by seven equally spaced inner knots over $[0, 1]$, where 0 and 1 indicate the opening (9:30 a.m.) and closing (15:00 p.m.) of the market, respectively. The contribution rates of variance of the largest six generated basis functions are 0.641, 0.164, 0.073, 0.045, 0.026 and 0.021, reaching 0.971 in total. Fig. 2 depicts the shapes of these six orthonormal basis functions. As illustrated by Fig. 2, the first three basis functions in the top panel show different patterns of volatility in IPR, while the next three in the bottom panel just differ in periodicity.

Then, we conduct a series of CompLMM, where the random effects contain all the four covariates and IPR is represented by the expansion coefficients for the first one to six orthonormal basis functions, as well as CompLM. Table 3 reports the related estimated coefficients of the fixed effects, and the results show the consistence across different numbers of basis functions used in both CompLMM and CompLM, but two methods are remarkably different in the estimated results. Following the CCR-based criteria on choosing the number of orthonormal basis functions, we get $K = 3$, 4 and 5 for the threshold of CCR as $\delta = 0.85$, 0.9 and 0.95, respectively. To avoid the under- and over-fitting of the observed data from IPR, we choose $\delta = 0.9$ along with $K = 4$ in the further discussion.

Next, we construct the theoretical confidence intervals/bands for the estimated results under the significance level of $\varrho = 0.05$. We also carry out the bootstrap procedure to numerically evaluate the significance of the estimated results. Specifically, in each sampling procedure, we randomly take with replacement 90% samples from the full data, and conduct models on these data. The estimated coefficients for the ilr coordinates of OIS are transformed back to the original five sentiments, and those for the expansion coefficients of IPR under the orthonormal basis functions are used to reconstruct the functional coefficient in a point-wise manner. We independently replicate the aforementioned sampling procedure 1000 times, and build up the bootstrap-based confidence intervals/bands bounded by the related $\varrho/2$ and $1 - \varrho/2$ quantiles. Table 4 reports the estimated results using the full data for the scalar and clr-transformed compositional coefficients and related confidence intervals obtained by both statistical inference and bootstrap methods, and Fig. 3 illustrates the curves and pie charts of the estimate functional and compositional coefficients and related confidence bands by two methods.

As shown in Table 4, the contribution of DV to the stock price is contrasting in two methods: positive in CompLMM and negative in CompLM. Two methods also differ in the estimated clr-transformed compositional coefficients for OIS, although they both consider "Joy" and "Fear" as two important sentiments in explaining DCP. In CompLMM, as shown in the right panel of Fig. 3, "Joy" has the largest positive influence on DCP, with "Fear" tied for the second; on the contrary, two sentiments exchange the order in CompLM. Since the increase in an inner part of a composition implies a general decrease in the others, it is hard to separately evaluate the influence of a specific part (Pawlowsky-Glahn et al., 2015). Moreover, the confidence intervals obtained by both statistical inference and bootstrap methods show consistent, and the range for the bootstrap method is generally larger than that for the statistical inferences. Under $\varrho = 0.05$, the scalar covariate DV shows the significant contribution to DCP in both models. In CompLMM, the marginal contributions of all the five sentiments in OIS pass the significance test, while three of them fail in CompLM. And the large standard deviation of residuals for CompLM (reaching $\widehat{\sigma} = 12.464$) indicates the bad fitting of CDP; by contrast, $\widehat{\sigma} = 1.553$ shows the reasonable results by CompLMM.

Table 3
Estimated coefficients of the fixed effects based on the full data under different numbers of orthonormal basis functions used.

Model	Intercept	DV	IPR						OIS				
			ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6	Anger	Disgust	Joy	Sadness	Fear
CompLM	19.303	−0.097	−1.274	2.447	−1.219	−28.857	7.075	−15.45	0.026	0.227	0.219	0.104	0.425
	19.429	−0.098	−1.269	2.438	−1.237	−28.838	7.077	–[a]	0.029	0.244	0.23	0.105	0.392
	19.261	−0.098	−1.241	2.481	−1.265	−28.912	–	–	0.027	0.251	0.25	0.096	0.376
	20.285	−0.099	−1.391	2.357	−1.122	–	–	–	0.025	0.191	0.242	0.194	0.349
	20.278	−0.099	−1.396	2.354	–	–	–	–	0.025	0.185	0.241	0.195	0.354
	20.314	−0.099	−1.411	–	–	–	–	–	0.029	0.178	0.231	0.193	0.369
CompLMM	13.837	0.056	−0.343	0.983	−0.965	−2.468	2.018	−0.261	0.041	0.076	0.585	0.055	0.243
	13.984	0.052	−0.387	1.032	−0.924	−2.585	1.82	–	0.04	0.077	0.569	0.06	0.253
	13.933	0.052	−0.399	1.038	−1.007	−2.477	–	–	0.039	0.077	0.58	0.059	0.246
	13.974	0.053	−0.405	1.032	−1.013	–	–	–	0.039	0.075	0.579	0.061	0.246
	13.981	0.053	−0.42	1.085	–	–	–	–	0.039	0.073	0.576	0.062	0.25
	13.968	0.053	−0.408	–	–	–	–	–	0.042	0.072	0.568	0.061	0.257

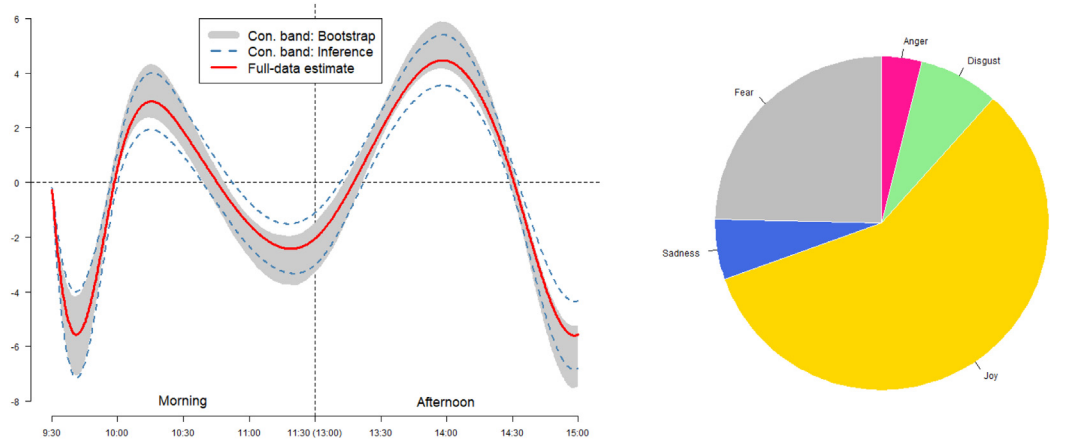
[a] indicates that the corresponding variable is not used.

Table 4
Estimated results using the full data for DV and OIS and related confidence intervals under $\varrho = 0.05$.

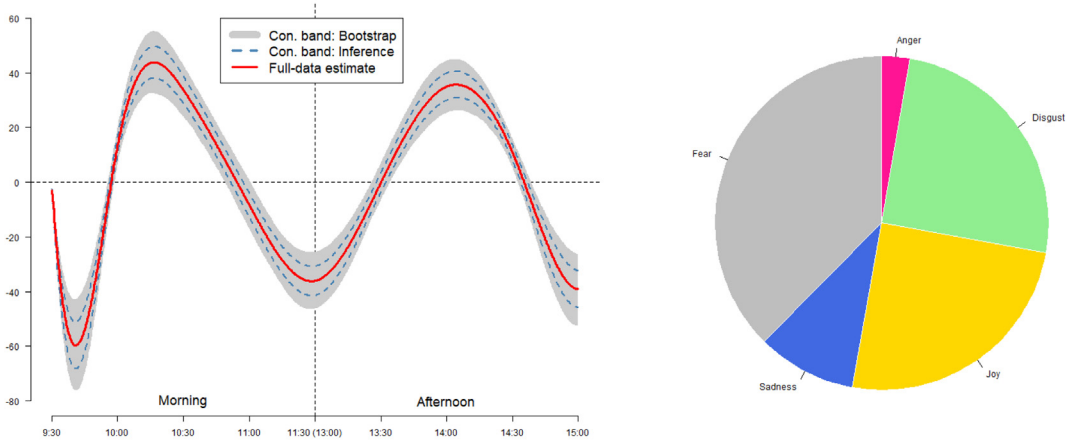
Model	Confidence interval		Intercept	DV	OIS[a]					$\widehat{\sigma}$
					Anger	Disgust	Joy	Sadness	Fear	
CompLM	Full-data		19.261	−0.098	−1.675	0.56	0.554	−0.4	0.961	12.464
	Inference	Upper	20.775	−0.094	−0.481	1.462	1.295	0.557	1.777	–
		Lower	17.748	−0.101	−2.87	−0.342	−0.187	−1.356	0.15	–
	Bootstrap	Upper	20.842	−0.089	−0.418	1.541	1.358	0.62	1.826	–
		Lower	17.591	−0.107	−2.942	−0.413	−0.175	−1.469	0.084	–
CompLMM	Full-data		13.933	0.052	−1.133	−0.441	1.575	−0.718	0.718	1.553
	Inference	Upper	15.507	0.065	−0.917	−0.183	1.833	−0.421	0.972	–
		Lower	12.36	0.039	−1.349	−0.699	1.316	−1.015	0.463	–
	Bootstrap	Upper	14.111	0.064	−0.849	−0.328	1.760	−0.659	0.814	–
		Lower	13.521	0.046	−1.125	−0.602	1.516	−0.956	0.531	–

[a] denotes the clr-transformed results for OIS.

As displayed in the left panel of Fig. 3, the curves of the estimated functional coefficient by two methods share almost the same shape, where it is indicated that the returns near 10:00–10:30 a.m. and 14:00 p.m. show relatively high marginal contributions to DCP. Two curves differ in scale, and the range of values in CompLM is 10 times larger than that in CompLMM, which may account for the bad performance of CompLM to a great extent. Moreover, the confidence band has a narrower



(a) CompLMM.



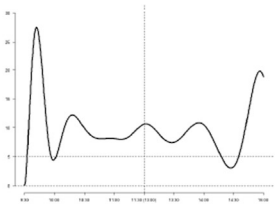
(b) CompLM.

Fig. 3. Curves and pie charts of the estimated functional and compositional coefficients for IPR and OIE, respectively. The red curve denotes the estimated result for IPR using the full data, the blue dotted curves and grey region indicate the confidence bands obtained by the statistical inference and bootstrap methods, respectively, and the vertical dotted line divides a trading day into the morning and afternoon.

Table 5
Estimated covariance matrix of the random effects for CompLMM in the real data study on Chinese stock market. The sub-columns ξ_j ($j = 1,2,\ldots,5$) denote the expansion coefficients under the orthonormal basis functions for IPR and w^*_{1k} ($k = 1,2,\ldots,4$) indicate the ilr coordinates of the compositional coefficients for OIE. The variances are highlighted in bold. The variance function and total covariance respectively for IPR and OIE are also plotted and reported.

	Intercept	DV	IPR				OIE			
			ξ_1	ξ_2	ξ_3	ξ_4	w^*_{11}	w^*_{12}	w^*_{13}	w^*_{14}
Intercept	**161.754**	0.33	-2.637	4.042	-0.7	4.523	9.172	-1.706	-4.601	-12.946
DV		**0.01**	0.017	-0.019	0.025	0.001	0.05	0.091	-0.077	0.123
ξ_1			**0.415**	-0.227	0.133	0.006	0.013	0.205	-0.225	0.668
ξ_2				**1.477**	-0.565	0.082	-0.187	-0.386	0.621	-1.394
ξ_3					**3.795**	-0.109	0.402	-0.099	-0.636	0.794
ξ_4						**4.223**	0.5	-0.562	-0.331	-0.729
w^*_{11}							**1.994**	-0.379	-2.45	1.702
w^*_{12}								**4.324**	-0.093	2.445
w^*_{13}									**3.72**	-3.772
w^*_{14}										**7.256**

Variance function for IPR



Total variance for OIE: 17.294

width by the statistical inferences than by the bootstrap method in CompLM; while in CompLMM, they have slightly differences in the peaks and troughs of the curve, and the confidence band by the statistical inferences shows less volatility than by the bootstrap method.

Finally, we focus on the estimated results for the random effects for CompLMM, as reported in Table 5. The large variance of the intercept (i.e., 161.754) indicate that the stocks involved have great differences in prices. The variance of the DV is relatively small (only 0.01), which implies that its influence has few changes across stocks; therefore, the DV could be regarded as an inessential factor in the random effects in this case. To describe the overall variation of the functional coefficient, we plot the variance function based on the covariance matrix of the expansion coefficients, where the point-wise variances during most trading hours hover around 10 in general, and the curve contains two peaks that correspond to the time before 10:00 a.m. and after 14:30 p.m. This result verifies that the past trends of the return from different stocks, especially at the beginning and ending time, have diversified influences on their prices in the future. Finally, we sum up the variances of the four ilr coordinates and obtain the related total variance of the compositional coefficients for OIE as 17.294. This result verifies that the indirect information such as OIE, although shared by all the stocks, could also enhance the performance of the regression model for the stock price in various ways.

In conclusion, the real data study on Chinese stock market illustrates the potential of the proposed CompLMM for longitudinal complex data from an application perspective. Introducing the random effects containing the scalar, functional and compositional covariates, our method measures the subject-specific characteristics of each stock and improves the performance of the regression on diversified types of variables from different sources. However, there remains some problems to be discussed from both theoretical and practical perspectives, such as a more exhaustive explanation of the functional or compositional coefficients and the choice of the direct and indirect indicators for Chinese stock market. These issues need to be addressed in future work.

6. Concluding remarks

This study investigates an LMM technique for longitudinal complex data named CompLMM, involving a scalar continuous response and complex data covariates with diversified characteristics. Through the random effects that describe the differences across individuals, CompLMM can extract further information from the residuals obtained by the existing linear model for complex data and then show a significant improvement in fitting responses. Following the linear framework of complex data modeling, CompLMM first consistently represents different types of variables such that the classic LMM can then be conducted to obtain the intermediate results and transform them back to have the related diversified features. This model also encourages a more comprehensive interpretation for the regression on complex data. Moreover, some theoretical properties are also presented that support the computational procedure of the parameter estimation. As illustrated by both the simulation studies and real data analysis, the proposed CompLMM succeeds in dealing with longitudinal complex data and efficiently estimating the parameters with more reliable response fitting.

We focus on the parameter estimation and general interpretation for the proposed CompLMM. Through the consistent numeric representation for multiple types of complex data, the multivariate normality on the random effects is promising to

be generalized to more flexible distributions such as the multivariate t distribution (Pinheiro et al., 2001) to realize a more reliable and efficient estimation. The maximum likelihood estimation based on a new distribution and related EM algorithm will be reconsidered and reformulated accordingly. Indeed, it is worthy of study in the future.

Next, the trade-off between the accuracy and interpretation of the proposed model also needs due consideration, to which many solutions for the traditional scalar covariates have been proposed. These statistical methods provide instructive strategies for selecting random effects with diversified characteristics, which faces great challenges in theory but deserves further research. The practical and empirical ways of determining the random effects also demand investigation. Moreover, many types of complex data, as discussed in Section 3.3, have the potential to be modeled under the proposed framework using suitable representations. The processing of these variables has been adopted by many studies, but some of the theoretical properties for this study need detailed checks in the future.

Finally, the statistical inferences on multiple types of complex data are also an important and challenging issue in regression. In this study, we investigate the statistical inferences under the normality assumption, however, some problems remain to be addressed. For example, the direct hypothesis test on the original compositional data faces great challenges due to the uncertain magnitude of scale in the closure operation. Although some empirical methods such as the bootstrap could partly offer the solutions to these problems, the theory of the related hypothesis tests for complex data should be developed. Moreover, the truncation of the basis functions approximately represents the infinite-dimensional functions as few expansion coefficients, but unavoidably leads to the extra bias in the error. The related asymptotic property for the truncated expansion coefficients in the proposed method needs further study.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgement

The authors are grateful to the Associate Editor and anonymous referees for their very constructive comments and suggestions which helped improve the paper greatly. This research was financially supported by the Natural Science Foundation of China (Nos. 71420107025, 11701023).

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B, 44*, 139–160.
Aitchison, J. (1986). *The statistical analysis of compositional data*. Springer.
Aitchison, J., & Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika, 71*, 323–330.
Billard, L., & Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association, 98*, 470–487.
Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics, 66*, 1069–1077.
Bruno, F., Greco, F., & Ventrucci, M. (2014). Spatio-temporal regression on compositional covariates: Modeling vegetation in a gypsum outcrop. *Environmental and Ecological Statistics, 22*, 445–463.
Bruno, F., Greco, F., & Ventrucci, M. (2016). Non-parametric regression on compositional covariates using Bayesian P-splines. *Statistical Methods and Applications, 25*, 75–88.
Chen, L., & Cao, H. (2017). Analysis of asynchronous longitudinal data with partially linear models. *Electronic Journal of Statistics, 11*, 1549–1569.
Chen, N. F., Roll, R., & Ross, S. A. (1986). Economic forces and the stock market. *Journal of Business, 59*, 383–403.
Duan, J., Liu, H., & Zeng, J. (2013). Posterior probability model for stock return prediction based on analyst's recommendation behavior. *Knowledge-Based Systems, 50*, 151–158.
Egozcue, J. J., & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology, 37*, 795–828.
Egozcue, J. J., Pawlowskyglahn, V., Mateu-Figueras, G., & Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology, 35*, 279–300.
Fan, Y., James, G. M., & Radchenko, P. (2015). Functional additive regression. *Annals of Statistics, 43*, 2296–2325.
Filzmoser, P., Hron, K., & Templ, M. (2018). *Applied compositional data analysis: With worked examples in R*. Springer.
Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. John Wiley & Sons.
Gertheiss, J., Goldsmith, J., Crainiceanu, C., & Greven, S. (2013). Longitudinal scalar-on-functions regression with application to tractography data. *Biostatistics, 14*, 447–461.
Goldsmith, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 61*, 453–469.
Hall, P., & Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society: Series B, 78*, 637–653.
Hall, P., & Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *Annals of Statistics, 35*, 70–91.
Härdle, W. K., Müller, M., Sperlich, S., & Werwatz, A. (2012). *Nonparametric and semiparametric models*. Springer Science & Business Media.
Hsiao, C. (2014). *Analysis of panel data*. Cambridge University Press.
Irpino, A., & Verde, R. (2015). Linear regression for numeric symbolic variables: A least squares approach based on wasserstein distance. *Advances in Data Analysis and Classification, 9*, 81–106.
Laird, N. M., Lange, N., & Stram, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *Journal of the American Statistical Association, 82*, 97–105.
Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics, 38*, 963–974.
Marzio, M. D., Panzera, A., & Venieri, C. (2015). Non-parametric regression for compositional data. *Statistical Modelling, 15*, 113–133.
Mateu-Figueras, G., Pawlowsky-Glahn, V., & Egozcue, J. J. (2013). The normal distribution in some constrained sample spaces. *Sort Statistics & Operations Research Transactions, 37*, 29–56.
Müller, H. G., & Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics, 33*, 774–805.
Pawlowsky-Glahn, V., Egozcue, J. J., & Tolosana-Delgado, R. (2015). *Modelling and analysis of compositional data*.

Pinheiro, J. C., Liu, C., & Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational & Graphical Statistics, 10*, 249–276.
Qiu, Z., Song, P. X. K., & Tan, M. (2010). Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics, 35*, 577–596.
Ramsay, J. O. (1982). When the data are functions. *Psychometrika, 47*, 379–396.
Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer.
Ramsay, J. O., & Silverman, B. W. (2007). *Applied functional data analysis: Methods and case studies*. Springer.
Ruan, Y., Durresi, A., & Alfantoukh, L. (2018). Using twitter trust network for stock market analysis. *Knowledge-Based Systems, 145*, 207–218.
Sun, Y., Han, A., Hong, Y., & Wang, S. (2018). Threshold autoregressive models for interval-valued time series data. *Journal of Econometrics, 206*, 414–446.
Wang, H., Huang, L., Shangguan, L., & Wang, S. (2015). Variable selection and estimation for regression models with compositional data predictors. In *International workshop on compositional data analysis*.
Wang, S., Huang, M., Wu, X., & Yao, W. (2016). Mixture of functional linear models and its application to CO2-GDP functional data. *Computational Statistics & Data Analysis, 97*, 1–15.
Wang, H., Lu, S., & Zhao, J. (2019a). Aggregating multiple types of complex data in stock market prediction: A model-independent framework. *Knowledge-Based Systems, 164*, 193–204.
Wang, Z., Wang, H., & Wang, S. (2019b). Linear mixed-effects model for multivariate compositional data. *Neurocomputing, 335*, 48–58.
Wei, Y., Wang, S., & Wang, H. (2017). Interval-valued data regression using partial linear model. *Journal of Statistical Computation and Simulation, 87*, 3175–3194.
Zhang, P., Qiu, Z., & Song, P. X. K. (2009). Robust transformation mixed-effects models for longitudinal continuous proportional data. *Canadian Journal of Statistics/La Revue Canadienne de Statistique, 37*, 266–281.
Zhou, Z., Xu, K., & Zhao, J. (2018). Tales of emotion and stock in China: Volatility, causality and prediction. *World Wide Web, 21*, 1093–1116.