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

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

PII: S2096-2320(19)30089-7

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

Reference: JMSE 14

To appear in: *Journal of Management Science and Engineering*

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

Authors:
Zhichao Wang[1], Huiwen Wang[1,2], Shanshan Wang[1,2], Shan Lu[3], Gilbert Saporta[4]

A_liation:
1 School of Economics and Management, Beihang University, Beijing 100191, China;
2 Beijing Advanced Innovation Center for Big Data and Brain Computing, Beijing 100191, China;
3 School of Statistics and Mathematics, Central University of Finance and Economics, Beijing 100081, China;
4 Applied Statistics, Conservatoire National des Arts et M_etiers, Paris 75141, France.

Correspondence:

Zhichao Wang: wangzc1415@buaa.edu.cn

Huiwen Wang: wanghw@vip.sina.com

Shanshan Wang: sswang@buaa.edu.cn

Shan Lu: lushan@buaa.edu.cn

Gilbert Saporta: gilbert.saporta@cnam.fr

* Corresponding author.
Correspondence to: School of Economics and Management, Beihang University, Beijing, China.
E-mail address: sswang@buaa.edu.cn (S.S. Wang).
Acknowledgement:
This research was _nancially supported by the Natural Science Foundation of China (Nos. 71420107025, 11701023).

Conflicts of onterest
The authors declare no conflict of interest.

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

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, 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 a traditional longitudinal framework. In this study, we investigate a 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 LMM can then be conducted on them, which generalizes the theoretical framework of the LMM to complex data analysis. A number of numerical experiments indicate the feasibility and effectiveness of the proposed model. We further illustrate its practical utility in a real data study of China's stock market and show that the proposed method can enhance the performance and interpretability of the regression for complex data with diversified characteristics.

Key Words: Longitudinal complex data; Linear mixed-effects model; Compositional data analysis; Functional data analysis; Stock market; Online investors' emotions

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 familiar single points, curves, and pie charts. Multiple types of data, referred to as *complex data*, 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, public online emotion from social media and intraday stock returns are improving the accuracy and interpretation of trend predictions in China's stock market (Wang et al., 2019a). In such a case, the return series are processed as continuous functions from opening to closing and investors' emotions are measured as compositions constituted by five types of emotions.

Complex data analysis involves various types of data ranging from classical nominal, ordinal, and ratio scalar variables to curve- and pie-like functional and compositional variables and even text data.

In this study, we focus on two emerging types, namely functions and compositions. 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 internal property of functions (i.e., the infinite dimension) causes great difficulty for functional modeling in both theory and practice. To describe functions over a bounded closed set (e.g., an interval), some equivalent representations such as basis function expansion and the reproducing kernel method 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 relative information (Aitchison, 1982, 1986). The defining features of compositions include the strict positive and constant sum of all the components inside (e.g., 1 for proportions and 100 for percentages), which is also problematic for most traditional statistical approaches. To eliminate these strong constraints, a family of logratio transformations has been proposed such as additive logratio, centered logratio (Aitchison, 1986), and isometric logratio (Egozcue et al., 2003), abbreviated to the well-known *alr*, *clr*, and *ilr* transformations, respectively. For further details on FDA and CDA, see Ramsay and Silverman (2005) and Pawlowsky-Glahn et al. (2015), respectively.

Numerous works have investigated regression for functional and compositional covariates 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 follow the additive model (Fan et al., 2015), mixture of linear models (Wang et al., 2016), and truncated linear model (Hall and Hooker, 2016). Meanwhile, Aitchison and Bacon-Shone (1984) initially proposed linear regression for compositional covariates, Marzio et al. (2015) presented the kernel-based compositional regression, and Bruno et al. (2014, 2016) investigated the spatiotemporal model and another nonparametric regression with Bayesian P-splines, respectively.

The aforementioned approaches focus on a specific type of complex data, with relatively few studies examining the multi-type situation. Wang et al. (2015) preliminarily developed a linear model for multiple types of complex data. Wang et al. (2019a) then extended its computational framework to generalized linear regression including scalar, functional, and compositional covariates. Their methods were performed under the independent and identically distributed (IID) assumption of errors, namely that all the samples of complex 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 introduced in Section 5, the closing prices of China's stock market were collected from hundreds of stocks over several months. The price-limiting mechanism of the market results in high correlation among observations of the same stock. In such a case, statistical models of complex data based on the IID assumption can be biased and lead to confusing results since such longitudinal features are ignored. Similar problems have been discussed separately for FDA (Goldsmith et al., 2012; Gertheiss et al., 2013; Chen and Cao, 2017) and CDA (Zhang et al., 2009; Qiu et al., 2010; Wang et al., 2019b), and few of the proposed solutions show the potential to integrate multiple types of complex data. Thus, developing a unified framework to model longitudinal complex data with diversified characteristics is necessary.

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 focus only on scalar variables. In this study, we investigate the 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 ($i = 1, 2, \cdots, N$); then, CompLMM is formulated as

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

for the j-th sample ($j = 1, 2, \cdots, n_i$). Here, y_{ij} denotes the scalar response; x_{ijk} (z_{ijk}), μ_{ijk} (ν_{ijk}), and $\boldsymbol{c}_{ijk}$ ($\boldsymbol{w}_{ijk}$) are constituted by part of the scalar, functional, and compositional covariates, with numbers 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; ε_{ij} is the random scalar error; and $(\cdot,\cdot)_a$ denotes the Aitchison inner product in CDA to be introduced in Section 2. In the paradigm of the 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 one.

In particular, when there are no random effects, Model (1) is categorized as the computational framework of complex data in Wang et al. (2019a) and this reduces to the IID-based linear model (Wang et al., 2015), say CompLM, as

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

Compared with CompLM, the introduction of random effects into Model (1) makes it possible to capture the subject-specific information of each individual on the basis of the common regression characteristics in the population. It also distinguishes the between- and within-subject variability of responses, which further improves the performance of linear regression on longitudinal complex data.

In this study, we estimate the parameters for Model (1). To consistently represent data with diversified characteristics, we first transform the functions and compositions inside using related basis expansions such that they are described equivalently to numeric coordinates. These processed data are available to conduct the LMM and obtain the intermediate result that can then be reconstructed to match the original diversified characteristics. Then, 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, and 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 describes a real data study on China's stock market to illustrate the effectiveness of the proposed method. Finally, some discussions and prospects are given in Section 6.

2 Preliminaries

We briefly introduce the basic ideas and mathematical techniques for FDA and CDA, including basis function expansion for functions and ilr transformation for compositions. These provide the theoretical and computational foundation for the proposed method. For simplification, 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 time. Basis function expansion is one of the most practical methods of describing the continuous characteristics of a function (Ramsay and Silverman, 2005). That is, a function is expressed as a linear combination of the given basis functions, which can be realized using ordinary least squares (OLS), penalized OLS, or regularized principal components (Hall and Horowitz, 2007). Without loss of generality, we adopt B-spline basis functions and perform the simple OLS-based expansion in this study.

Specifically, given a group of basis functions $\{\phi_j\}_{j=1}^{\infty}$ over an interval $\mathcal{I}$, any square-integrable function, say $\mu \in \mathcal{L}_2$, can be formulated as $\mu = \sum_j u_j \phi_j$ with an infinite series of expansion coefficients u_j. When n samples, say o_i at time $t_i \in \mathcal{I}$ $(i = 1, 2, \cdots, n)$, are observed from μ, they are assumed subject to $o_i = \mu(t_i) + \epsilon_i$ with white noise ϵ_i. Then, the expansion of μ leads to $\boldsymbol{o} = \boldsymbol{\Phi}\boldsymbol{u} + \boldsymbol{\epsilon}$, where

$\boldsymbol{o} = (o_1, o_2, \cdots, o_n)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)'$ in $\mathbb{R}^n$, $\boldsymbol{\Phi} = \big(\boldsymbol{\phi}(t_1), \boldsymbol{\phi}(t_2), \cdots, \boldsymbol{\phi}(t_n)\big)' \in \mathbb{R}^{n\times K}$ with $\boldsymbol{\phi}(t_i) = \big(\phi_1(t_i), \phi_2(t_i), \cdots, \phi_K(t_i)\big)' \in \mathbb{R}^K$, and $\boldsymbol{u} = (u_1, u_2, \cdots, u_K)' \in \mathbb{R}^K$. In practice, the number of basis functions and related expansion coefficients K are limited below n because of the finite size of observations. Thus, the OLS estimation results in the truncated expansion coefficients of $\boldsymbol{u}$:

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

The expansion coefficients greatly concentrate the features of the original function. Typically, the image of μ can be described explicitly in a point-wise manner as

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

where μ is determined completely by $\boldsymbol{u}$ along with the known basis functions $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_K)'$. The expectation of μ is also associated with that of $\boldsymbol{u}$, and the variance function for μ, denoted by $\mathcal{K}_\mu$, can be formulated as

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

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 its expansion coefficients $\boldsymbol{\lambda}$, can be written as

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

To compute the integral in $\boldsymbol{W}$ numerically, we uniformly sample from $\mathcal{I}$, say $\{\tau_1, \tau_2, \cdots, \tau_T\}$ with T points and approximate it as $\boldsymbol{W} = T^{-1}\sum_{i=1}^{T}\boldsymbol{\phi}(\tau_i)'\boldsymbol{\phi}(\tau_i)$. $\boldsymbol{W}$ can also be adopted to express the overall difference between two functions as

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

with $\hat{\beta} \in \mathcal{L}_2$ and its expansion coefficients $\widehat{\boldsymbol{\lambda}}$. These properties of basis function expansion make it possible to equivalently represent infinite-dimensional functions as relatively few numeric variables.

2.2 CDA

In CDA, the attraction of a multivariate vector is the relative magnitude, instead of the absolute one, among all the components inside. Working with this scale invariance property, any composition with D inner parts, say $\boldsymbol{c}$, can be expressed as $\boldsymbol{c} = (c_1, c_2, \cdots, c_D)'$ subject to $c_i > 0$ $(i = 1, 2, \cdots, D)$ and $\boldsymbol{c}'\mathbf{1}_D = 1$ with $\mathbf{1}_D$ constituted by 1 in $\mathbb{R}^D$. All such D-part compositions consist of the D-dimensional simplex space denoted by S^D.

To remove the constraints of compositions, Egozcue et al. (2003) proposed the ilr transformation via the simplicial orthonormal basis. In this study, we follow Egozcue and Pawlowsky-Glahn (2005) and represent any composition as its specified coordinates. Take $\boldsymbol{c}$, for example, $\mathrm{ilr}(\boldsymbol{c}) = \boldsymbol{c}^* = (c_1^*, c_2^*, \cdots, c_{D-1}^*)'$, where

$$c_i^* = \frac{1}{\sqrt{(D-i+1)(D-i)}} \sum_{j=1}^{D-i} \log c_j - \sqrt{\frac{D-i}{D-i+1}} \log c_{D-i+1} \quad (i = 1, 2, \cdots, D-1). \tag{7}$$

These coordinates contain all the relative information on $\boldsymbol{c}$; therefore, they can be used to reconstruct the original composition. That is, $\boldsymbol{c} = \mathrm{ilr}^{-1}(\boldsymbol{c}^*) = \mathcal{C}\big(\exp(\boldsymbol{\omega})\big)$, where $\mathcal{C}(\cdot)$ denotes the closure operation that scales a vector with positive components proportionally such that it conforms to the constraints of compositions, and $\exp(\boldsymbol{\omega}) = (\exp\omega_1, \exp\omega_2, \cdots, \exp\omega_D)'$ with

$$\omega_i = \sum_{j=0}^{D-i} \frac{c_j^*}{\sqrt{(D-j+1)(D-j)}} - \sqrt{\frac{i-1}{i}} c_{D-i+1}^* \quad (i = 1, 2, \cdots, D) \tag{8}$$

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 $\mathrm{ilr}^{-1}(\boldsymbol{c}^*) = \mathcal{C}\big(\exp(\boldsymbol{\Psi}'\boldsymbol{c}^*)\big)$, where $\log(\boldsymbol{c}) = (\log c_1, \log c_2, \cdots, \log c_D)'$. Specifically, $\boldsymbol{\Psi}$ associated with (7) and (8) is constituted by the elements ψ_{ij} as $\psi_{ij} = \sqrt{\frac{D-i}{D-i+1}}\rho_{ij}$ for $i = 1, 2, \cdots, N$ and $j = 1, 2, \cdots, n_i$, 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 Euclidian spaces, the ilr transformation facilitates the computation of the Aitchison geometry. For example, the Aitchison inner product, denoted by $(\cdot,\cdot)_a$, can be easily expressed as

$$(\boldsymbol{c}, \boldsymbol{\gamma})_a = \mathrm{ilr}(\boldsymbol{c})'\mathrm{ilr}(\boldsymbol{\gamma}) \quad (\boldsymbol{\gamma} \in S^D). \tag{9}$$

The related norm and distance, denoted by $\|\cdot\|_a$ and $d_a(\cdot,\cdot)$, then follow respectively as

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

with $\widehat{\boldsymbol{\gamma}} \in S^D$ and its ilr coordinates $\widehat{\boldsymbol{\gamma}}^*$. Moreover, the total variance of $\boldsymbol{c}$, denoted by $\mathrm{totVar}(\boldsymbol{c})$, can be decomposed as

$$\mathrm{totVar}(\boldsymbol{c}) = \sum_{i=1}^{D-1} \mathrm{Var}(c_i^*) \tag{10}$$

using the ilr coordinates. Since $\boldsymbol{\Psi}'\boldsymbol{\Psi}$ is identically equal to $\boldsymbol{I}_D - \boldsymbol{1}_D\boldsymbol{1}_D'/D$, where $\boldsymbol{I}_D$ denotes the D-dimensional unit matrix (Pawlowsky-Glahn et al., 2015), the specified ilr transformation here does not affect those results above. From these properties of the ilr transformation, we can substitute the ilr coordinates with no constraints for the compositional covariates in most statistical models.

3 LMM for complex data

In this section, we investigate the LMM for longitudinal complex data with diversified characteristics. The approach used to aggregate multiple types of complex data along with their properties and some issues in practice are also discussed.

3.1 Model

To uniformly represent complex data with different characteristics, we apply the B-spline expansion and ilr transformation to Model (1). Thus, the model can be formulated using (5) and (9) as

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

where $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \cdots, x_{ij,p_x})' \in \mathbb{R}^{p_x}$ with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_{p_x})'$ and $\boldsymbol{z}_{ij} = (z_{ij1}, z_{ij2}, \cdots, z_{ij,q_z})' \in \mathbb{R}^{q_z}$ with $\boldsymbol{a}_i = (a_{i1}, a_{i2}, \cdots, a_{i,q_z})'$; $\boldsymbol{u}_{ijk}$, $\boldsymbol{v}_{ijk}$, $\boldsymbol{\lambda}_k$ and $\boldsymbol{\theta}_{ik}$ denote the expansion coefficients of μ_{ijk}, ν_{ijk}, β_k and b_{ik}, respectively, with a common dimension K; 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

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

where $\boldsymbol{u}_{ij} = (\boldsymbol{u}_{ij1}; \boldsymbol{u}_{ij2}; \cdots; \boldsymbol{u}_{ij,p_\mu}) \in \mathbb{R}^{Kp_\mu}$ and $\boldsymbol{v}_{ij} = (\boldsymbol{v}_{ij1}; \boldsymbol{v}_{ij2}; \cdots; \boldsymbol{v}_{ij,q_\nu}) \in \mathbb{R}^{Kq_\nu}$, along with $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1; \boldsymbol{\lambda}_2; \cdots; \boldsymbol{\lambda}_{p_\mu})$ and $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}; \boldsymbol{\theta}_{i2}; \cdots; \boldsymbol{\theta}_{i,q_\mu})$, respectively; $\boldsymbol{W}_{p_\mu}$ ($\boldsymbol{W}_{q_\nu}$) denotes the blocked diagonal matrix consisting of p_μ (q_ν) matrices $\boldsymbol{W}$; and $\boldsymbol{c}^*_{ij} = (\boldsymbol{c}^*_{ij1}; \boldsymbol{c}^*_{ij2}; \cdots; \boldsymbol{c}^*_{ij,p_c}) \in \mathbb{R}^{(D-1)p_c}$ and $\boldsymbol{w}^*_{ij} = (\boldsymbol{w}^*_{ij1}; \boldsymbol{w}^*_{ij2}; \cdots; \boldsymbol{w}^*_{ij,q_w}) \in \mathbb{R}^{(D-1)q_w}$, along with $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^*_1; \boldsymbol{\gamma}^*_2; \cdots; \boldsymbol{\gamma}^*_{p_c})$ and $\boldsymbol{r}^*_i = (\boldsymbol{r}^*_{i1}; \boldsymbol{r}^*_{i2}; \cdots; \boldsymbol{r}^*_{i,q_w})$, respectively.

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

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

where $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \cdots, y_{i,n_i})'$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{i,n_i})'$ in $\mathbb{R}^{n_i}$; specifically, the fixed effects involve $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{i,n_i})' \in \mathbb{R}^{n_i \times p_x}$, $\boldsymbol{u}_i = (\boldsymbol{u}_{i1}, \boldsymbol{u}_{i2}, \cdots, \boldsymbol{u}_{i,n_i})' \in \mathbb{R}^{n_i \times p_\nu}$ and $\boldsymbol{c}^*_i = (\boldsymbol{c}^*_{i1}, \boldsymbol{c}^*_{i2}, \cdots, \boldsymbol{c}^*_{i,n_i})' \in \mathbb{R}^{n_i \times p_c}$, and the random effects involve $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \boldsymbol{z}_{i2}, \cdots, \boldsymbol{z}_{i,n_i})' \in \mathbb{R}_{n_i \times q_z}$, $\boldsymbol{v}_i = (\boldsymbol{v}_{i1}, \boldsymbol{v}_{i2}, \cdots, \boldsymbol{v}_{i,n_i})' \in \mathbb{R}^{n_i \times q_\nu}$ and $\boldsymbol{w}^*_i = (\boldsymbol{w}^*_{i1}, \boldsymbol{w}^*_{i2}, \cdots, \boldsymbol{w}^*_{i,n_i})' \in \mathbb{R}^{n_i \times q_w}$. To coincide with the paradigm of the LMM for the scalar variables 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 the dimensions of $p = p_x + Kp_\mu + (D-1)p_c$ and $q = q_z + Kq_\nu + (D-1)q_w$, respectively. Here, $\boldsymbol{\varpi}$ contains the common characteristics shared by the entire population and $\boldsymbol{\pi}_i$ shows the specific ones of the i-th individual. Moreover, $\boldsymbol{\varepsilon}_i$ is assumed to obey the normal distribution in $\mathbb{R}^{n_i}$, namely $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\boldsymbol{0}_{n_i}, \sigma^2\boldsymbol{I}_{n_i})$, where $\boldsymbol{0}_{n_i}$ 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 in $\mathbb{R}^q$, namely $\boldsymbol{\pi}_i \sim \mathcal{N}(\boldsymbol{0}_q, \boldsymbol{G})$, where $\boldsymbol{G}$ is positively defined 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 these aforementioned assumptions, the estimate of $\boldsymbol{\Theta}$, denoted by $\widehat{\boldsymbol{\Theta}} = \{\widehat{\boldsymbol{\varpi}}, \widehat{\boldsymbol{G}}, \hat{\sigma}^2\}$, can be derived using the expectation maximum (EM) algorithm. Then, the fitted response, for example, $\hat{y}_{ij}$

in Model (11), can be expressed as

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

or $\hat{y}_{ij} = \boldsymbol{x}'_{ij}\widehat{\boldsymbol{\alpha}} + \boldsymbol{u}'_{ij}\boldsymbol{W}_{p_\mu}\widehat{\boldsymbol{\lambda}} + (\boldsymbol{c}^*_{ij})'\widehat{\boldsymbol{\gamma}}^*$ for the reduced CompLM, where $\widehat{\boldsymbol{\varpi}}$ consists of $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\lambda}}$ and $\widehat{\boldsymbol{\gamma}}^*$, and $\widehat{\boldsymbol{a}}_i$, $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{r}}^*_i$ can be obtained from $\widehat{\boldsymbol{\Theta}}$. Before we develop the estimation procedure, we discuss the relationship between the original and reconstructed models (i.e., Models (1) and (12)) in the following remarks.

Remark 1. The assumption of the independence of the random effects coefficients and errors is important for the theory of the LMM. We put this assumption on the reconstructed unified numeric variables, namely $\boldsymbol{\theta}_i$ and $\boldsymbol{r}^*_i$, in Model (12), which implies the independence of the original coefficients with diversified characteristics (e.g., b_{ik} and $\boldsymbol{r}_{ik}$) and the scalar errors in Model (1). The covariance between b_{ik} with functional characteristics and ε_{ij} is defined as point-wise (Ramsay and Silverman, 2005), namely $\mathrm{Cov}(b_{ik}(t), \varepsilon_{ij})$ for any $t \in \mathcal{I}$. Under the given basis functions $\boldsymbol{\phi}$, the expectations of the related expansion coefficients are equal to zero; then, we have

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

From (13), the independence assumption on the expansion coefficients, namely $\mathrm{Cov}(\boldsymbol{\theta}_{ik}, \varepsilon_{ij}) = \mathbf{0}_K$, is sufficient to that on the original overall function. On the contrary, 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., 2019b).

Remark 2. Another issue for the theory of the LMM follows the covariance matrix of the random effects coefficients. For functional variables, this refers to the covariance function, say $\mathcal{K}_{b_{ik},b_{ik'}}(s,t)$ for b_{ik} and $b_{ik'}$ with the expansion coefficients $\boldsymbol{\theta}_{ik'}$ at times s and t. Similar to (4), it can be formulated as

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

Specifically, it reduces to $\mathcal{K}_{b_{ik}}(s,t) = \boldsymbol{\phi}'(s)\mathrm{Var}(\boldsymbol{\theta}_{ik})\boldsymbol{\phi}(t)$ when $k' = k$. For compositional variables, say $\boldsymbol{r}_{ik}$ and $\boldsymbol{r}_{ik'}$ with the ilr coordinates $\boldsymbol{r}^*_{ik'}$, the covariance matrix, as Mateu-Figueras et al. (2013) suggested, can be naturally defined as $\mathrm{Cov}(\boldsymbol{r}_{ik}, \boldsymbol{r}_{ik'}) = \mathrm{Cov}(\boldsymbol{r}^*_{ik}, \boldsymbol{r}^*_{ik'})$. Then, 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

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

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 reconstructed model concentrates the covariance structure of multiple types of complex data.

3.2 Parameter estimation

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

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

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

where $\mathbb{X}_i = (\boldsymbol{x}_i, \boldsymbol{u}_i\boldsymbol{W}_\mu, \boldsymbol{c}_i^*) \in \mathbb{R}^{n_i\times p}$ for $i = 1, 2, \cdots, N$ and $M = \sum_{i=1}^{N} n_i$. Here, $\widehat{\boldsymbol{\varpi}}_{ols}$ and $\hat{\sigma}^2_{ols}$ also indicate the consistent estimates within the computational framework of Wang et al. (2019a) for linear regression since both studies imply the same model expression as CompLM.

In general, the estimation procedure for Model (12) can be implemented using the EM algorithm (Laird et al., 1987). Specifically, given a pair of estimates $\widehat{\boldsymbol{G}}^{(\omega)}$ and $(\hat{\sigma}^{(\omega)})^2$, where the superscript ω indicates the iteration and $\omega = 0$ denotes the initial values, $\widehat{\boldsymbol{\varpi}}^{(\omega)}$ is formulated as

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

with

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

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

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

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

where

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

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

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

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

$$\max\Big\{\big\|\widehat{\boldsymbol{\varpi}}^{(\omega)} - \widehat{\boldsymbol{\varpi}}^{(\omega-1)}\big\|_\infty, \big|(\hat{\sigma}^{(\omega)})^2 - (\hat{\sigma}^{(\omega-1)})^2\big|\Big\} < \delta, \tag{22}$$

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 $\widehat{\boldsymbol{\varpi}}$, denoted by $\widehat{\boldsymbol{\varpi}}^{(0)}$, is set to be $\widehat{\boldsymbol{\varpi}}_{ols}$, and those of the other parameters can then be computed from $\widehat{\boldsymbol{\varpi}}^{(0)}$ as

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

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

where $L = M - (N-1)q - p$ and $\widehat{\boldsymbol{\pi}}^{(0)} = (\mathbb{Z}_i'\mathbb{Z}_i)^{-1}\mathbb{Z}_i'(\boldsymbol{y}_i - \mathbb{X}_i\widehat{\boldsymbol{\varpi}}^{(0)})$. The aforementioned initialization for the estimation procedure begins with the reduced OLS-based linear regression (i.e., CompLM) and further abstracts the subject-specific information from the covariance structure of errors. Again, it verifies that the proposed CompLMM improves the performance of the final regression for longitudinal complex data compared with CompLM.

Remark 3. The proposed parameter estimation for CompLMM using 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 effects, namely no z_{ijk}, ν_{ijk}, and $\boldsymbol{w}_{ijk}$ in Model (1), $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)}$ reduces to become proportional to $\boldsymbol{I}_{n_i}$ with no $\mathbb{Z}_i$ involved in (17), implying that $\widehat{\boldsymbol{\varpi}}^{(\omega)}$ in (16) equals $\widehat{\boldsymbol{\varpi}}_{ols}$ in (14) for any ω. A similar conclusion on (19) can also be drawn, namely that $(\hat{\sigma}^{(\omega+1)})^2 \equiv \hat{\sigma}_{ols}^2$ since

$$\widehat{\boldsymbol{e}}_i^{(\omega)} = \boldsymbol{y}_i - \mathbb{X}_i\widehat{\boldsymbol{\varpi}}^{(\omega)} \quad \text{and} \quad (\hat{\sigma}^{(\omega)})^2(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{y}_i}^{(\omega)})^{-1} = \boldsymbol{I}_{n_i}$$

here. We conclude from these results that CompLM works exactly as the pooled method for longitudinal complex data.

In summary, Algorithm 1 presents the computational procedure of the proposed method for longitudinal complex data.

Algorithm 1 Computational procedure for CompLMM

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

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

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

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

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 (7);

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

4: **repeat**

5: Compute the intermediates $\widehat{\boldsymbol{\pi}}_i^{(\omega+1)}$ and $\widehat{\boldsymbol{e}}_i^{(\omega+1)}$ $(i = 1, 2, \cdots, N)$ 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 $\widehat{\boldsymbol{\Sigma}}^{(\omega+1)}_{\boldsymbol{y}_i}$ $(i = 1, 2, \cdots, N)$ 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**

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

3.3 Some issues

In this study, we exemplify the proposed framework for longitudinal complex data using three types of covariates: scalar, function, and composition. This framework is actually available for more types of variables 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, where the order relation is ignored because of the continuous response. For unstructured text data, we can summarize these into a series of compositions associated with the frequencies of topics or positions, similar to the measurement of investors' emotions by Zhou et al. (2017), and therefore analyze text data under the proposed framework.

The key technique for formulating CompLMM is to find a suitable representation of a specific type of complex data and related consistent algebraic system, such as the dummy variable expression for categorical, nominal, and ordinal covariates, 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 combined into the proposed framework.

For example, in symbolic data analysis, an interval-valued variable has special binary representations such as "Lower-Upper" and "Center-Radius" (Billard and 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 other symbolic variables in symbolic data analysis, histograms, and distribution functions, with more complicated characteristics based on the Wasserstein distance (Irpino and Verde, 2015), and consider the related random effects to extend them to the proposed framework. Finally, some theoretical properties for the random effects associated with diversified variables, such as Remarks 1–3, remain to be checked, which need further research in the future.

Next, the introduction of random effects promotes the performance of linear regression for longitudinal complex data, while the complexity of random effects leads 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 suitable selection of random effects. As an important issue for the 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 Numerical experiment

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 percentage error (APE) for compositions. These are respectively defined in Model (1) as

$$\text{SRE} = \sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_{ij})^2 / \sum_{i=1}^{N}\sum_{j=1}^{n_i} y_{ij}^2,$$
$$\text{ISE}(\hat{\beta}_k) = d^2_{\mathcal{L}^2}(\beta_k, \hat{\beta}_k) \quad (k = 1, 2, \cdots, p_\mu),$$
$$\text{APE}(\widehat{\boldsymbol{\gamma}}_k) = d_a(\boldsymbol{\gamma}_k, \widehat{\boldsymbol{\gamma}}_k)/\|\boldsymbol{\gamma}_k\|_a \times 100\% \quad (k = 1, 2, \cdots, p_c),$$

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

We generate the data from Model (1) as

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

where seven three-order B-spline basis functions defined by four equally spaced interior knots over $[0, 1]$, say $\boldsymbol{\phi} = \{\phi_1, \phi_2, \cdots, \phi_7\}$, and the ilr coordinates from (7) and (8) are adopted. The detailed parameter settings 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 $\boldsymbol{\phi}$, 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 $\{t_1, t_2, \cdots, t_n\}$ from the linear combinations of $\boldsymbol{\phi}$ with measurement errors, that is,

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

where the expansion coefficients of both functions are sampled from $\mathcal{N}(\mathbf{0}_7, \boldsymbol{I}_7)$, 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 $\mathcal{N}_S(\mathbf{0}_2, \boldsymbol{I}_2)$ (Mateu-Figueras et al., 2013), with the compositional coefficients $\boldsymbol{\gamma}_1 = (0.8, 0.1, 0.1)'$ and $\boldsymbol{\gamma}_2 = (0.2, 0.2, 0.6)'$ 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 $\boldsymbol{\phi}$ and ilr coordinates, say $\boldsymbol{\theta}_{i1}$ and $\boldsymbol{r}^*_{i1}$, respectively. The parameters involved, namely $\boldsymbol{\pi}_i = (a_{i0}; \boldsymbol{\theta}_{i1}; \boldsymbol{r}^*_{i1})$, are then jointly generated from $\mathcal{N}(\mathbf{0}_{10}, \boldsymbol{G})$, where $\boldsymbol{G}$ is blocked diagonal, namely $\boldsymbol{G} = \text{diag}(9, \boldsymbol{G}_{\theta^*}, 0.5\boldsymbol{I}_4, \boldsymbol{G}_r)$ 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, 1.5, or 3 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: $(N, n_i) = (100, 60)$, $(300, 60)$, and $(300, 90)$. For each case, we independently replicate the simulation 500 times and conduct the proposed CompLMM as well as the baseline CompLM to provide a comparison. Table 1 summarizes the general performances of the two models for the SNRs across the sample sizes. Table 2 reports the estimated results for the functional and compositional coefficients of the two models with $(N, n_i) = (100, 60)$, and those with the other settings of (N, n_i) are reported in the Appendix. Moreover, Fig. 1 visualizes the specific curves of the estimated functional coefficients in randomly selected replications.

As shown in Table 1, the estimated coefficients for the scalar covariates (including the intercept) obtained from both CompLMM and CompLM on average approximate the related ideal values, while those from CompLMM are more stable with lower standard deviations than the baseline: 0.013 (CompLMM) vs. 0.105 (CompLM) for $\hat{\alpha}_1$ with $(N, n_i) = (100, 60)$ and $\sigma = 0.5$. For the functions, CompLMM sharply improves the estimation efficiency of the function-type coefficients, for which both the means and the standard deviations of the ISE are multiple times lower than those from CompLM: 0.03 and 0.021 (CompLMM) vs. 1.165 and 0.953 (CompLM). Moreover, the two methods perform the same for the compositions, where the ISE values for CompLMM are slightly more stable than those for CompLM. Finally, CompLMM estimates σ well for the SNRs, whereas the estimates of CompLM far exceed the corresponding true values. When noise is extremely large (i.e., $\sigma = 3$), the related estimate may be biased in a poor sample (e.g., $(N, n_i) = (100, 60)$), whereas such bias is mitigated if the sample is sufficiently large. Moreover, CompLM almost fails to fit the responses, since the values of the SRE are on average above 0.4 when $\sigma \leq 1.5$; by contrast, for CompLMM, the average values of the SRE are significantly low and close to 0.

The aforementioned conclusions on the estimated parameters for the functional and compositional covariates are confirmed by the expansion coefficients for the functions and detailed estimates for the compositions, as shown in Table 2. Specifically, the average estimates of the functions and compositions using both CompLMM and CompLM approach the related ideal values for the SNRs. However, the relatively high standard deviations for the expansion coefficients from CompLM lead to nonsignificant regression results. For example, the test statistic for $\hat{\lambda}_{11}$ with $\sigma = 0.5$, simply measured by 2.946/3.154, is less than the threshold value at the significance level of 0.05 (or even larger), which implies that we cannot reject the null hypothesis $H_0 : \lambda_{11} = 0$. By contrast, the same test statistic from CompLMM, similarly measured by 2.962/0.485, is more than the threshold value at the 0.05 level (or even smaller). For the compositions, owing to the limited magnitude of the proportions inside, the two methods show

Table 1: Means and standard derivations (in brackets) of the three measures and estimated parameters for the scalar covariates and errors. The ideal values of the intercept and $\hat{\alpha}_1$ are 2 and 5, respectively and those of $\hat{\sigma}^2$ coincide with the true setting of σ.

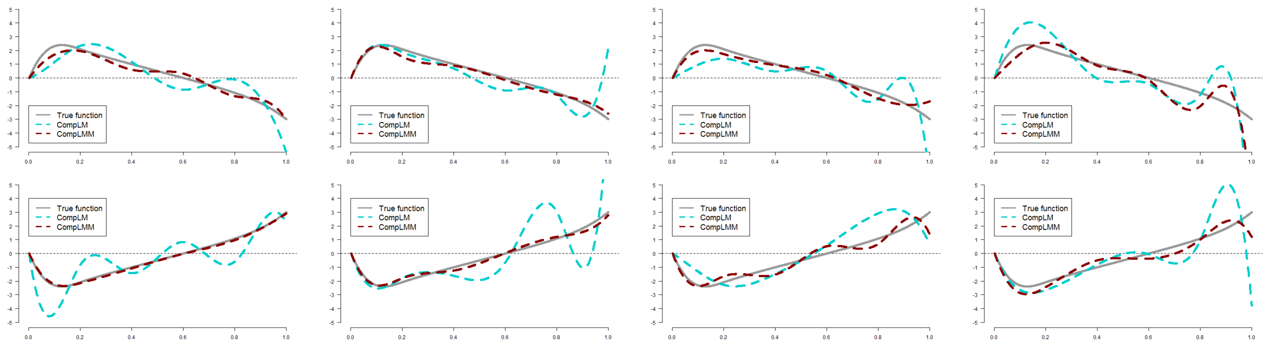
(N, n_i)	Model	Scalar		Functional		Compositional		$\hat{\sigma}^2$	SRE
		Intercept	$\hat{\alpha}_1$	ISE($\hat{\beta}_1$)	ISE($\hat{\beta}_2$)	APE($\widehat{\gamma}_1$)	APE($\widehat{\gamma}_2$)		
				$\sigma = 0.5$					
$(100, 60)$	CompLMM	1.97 (0.292)	5 (0.013)	0.03 (0.021)	0.016 (0.012)	6.185 (0.269)	0.448 (0.026)	0.263 (0.02)	0.016 (0.011)
	CompLM	1.971 (0.297)	4.996 (0.105)	1.165 (0.953)	1.194 (0.819)	6.177 (0.287)	0.447 (0.032)	22.314 (2.176)	0.425 (0.032)
$(300, 60)$	CompLMM	2.021 (0.176)	5 (0.007)	0.009 (0.007)	0.005 (0.004)	6.227 (0.164)	0.452 (0.016)	0.25 (0.003)	0.009 (0.003)
	CompLM	2.02 (0.177)	4.994 (0.06)	0.416 (0.299)	0.352 (0.244)	6.23 (0.168)	0.454 (0.018)	22.284 (1.178)	0.427 (0.019)
$(300, 90)$	CompLMM	2.007 (0.168)	5 (0.005)	0.007 (0.006)	0.003 (0.002)	6.238 (0.172)	0.453 (0.017)	0.25 (0.002)	0.01 (0.004)
	CompLM	2.005 (0.17)	5.002 (0.054)	0.275 (0.217)	0.243 (0.177)	6.24 (0.176)	0.454 (0.019)	22.288 (1.332)	0.427 (0.021)
				$\sigma = 1$					
$(100, 60)$	CompLMM	1.97 (0.293)	4.999 (0.024)	0.078 (0.054)	0.065 (0.049)	6.184 (0.027)	0.448 (0.026)	1.067 (0.08)	0.035 (0.011)
	CompLM	1.972 (0.298)	4.996 (0.107)	1.204 (0.977)	1.234 (0.855)	6.177 (0.287)	0.447 (0.032)	23.062 (2.179)	0.417 (0.032)
$(300, 60)$	CompLMM	2.021 (0.176)	5 (0.013)	0.023 (0.016)	0.02 (0.014)	6.227 (0.164)	0.452 (0.016)	0.998 (0.012)	0.027 (0.003)
	CompLM	2.02 (0.177)	4.994 (0.061)	0.433 (0.313)	0.364 (0.249)	6.23 (0.168)	0.454 (0.018)	23.033 (1.179)	0.419 (0.018)
$(300, 90)$	CompLMM	2.007 (0.168)	5 (0.011)	0.016 (0.011)	0.013 (0.01)	6.238 (0.172)	0.453 (0.017)	0.999 (0.009)	0.028 (0.004)
	CompLM	2.005 (0.17)	5.002 (0.055)	0.281 (0.22)	0.252 (0.183)	6.239 (0.176)	0.454 (0.019)	23.038 (1.332)	0.419 (0.02)
				$\sigma = 1.5$					
$(100, 60)$	CompLMM	1.971 (0.294)	4.998 (0.038)	0.162 (0.112)	0.147 (0.109)	6.183 (0.27)	0.448 (0.027)	2.436 (0.197)	0.066 (0.012)
	CompLM	1.972 (0.298)	4.995 (0.11)	1.27 (1.018)	1.3 (0.912)	6.176 (0.287)	0.447 (0.032)	24.309 (2.183)	0.404 (0.031)
$(300, 60)$	CompLMM	2.021 (0.176)	5 (0.02)	0.046 (0.033)	0.044 (0.032)	6.227 (0.164)	0.452 (0.016)	2.251 (0.032)	0.055 (0.003)
	CompLM	2.02 (0.177)	4.994 (0.063)	0.459 (0.333)	0.385 (0.261)	6.23 (0.168)	0.454 (0.018)	24.281 (1.18)	0.406 (0.018)
$(300, 90)$	CompLMM	2.007 (0.168)	5.001 (0.016)	0.031 (0.022)	0.029 (0.022)	6.238 (0.172)	0.453 (0.017)	2.249 (0.02)	0.056 (0.004)
	CompLM	2.005 (0.17)	5.002 (0.056)	0.292 (0.227)	0.268 (0.194)	6.239 (0.176)	0.454 (0.019)	24.288 (1.333)	0.406 (0.02)
				$\sigma = 3$					
$(100, 60)$	CompLMM	1.973 (0.297)	4.995 (0.076)	0.605 (0.416)	0.579 (0.414)	6.181 (0.273)	0.448 (0.028)	9.882 (0.826)	0.206 (0.022)
	CompLM	1.973 (0.302)	4.994 (0.125)	1.629 (1.26)	1.661 (1.195)	6.174 (0.289)	0.447 (0.033)	31.039 (2.212)	0.347 (0.027)
$(300, 60)$	CompLMM	2.021 (0.177)	5 (0.04)	0.169 (0.128)	0.174 (0.126)	6.227 (0.165)	0.452 (0.017)	9.052 (0.159)	0.183 (0.007)
	CompLM	2.02 (0.178)	4.994 (0.071)	0.59 (0.427)	0.504 (0.341)	6.23 (0.169)	0.454 (0.019)	31.022 (1.193)	0.348 (0.016)
$(300, 90)$	CompLMM	2.008 (0.169)	5.001 (0.031)	0.111 (0.084)	0.114 (0.087)	6.237 (0.173)	0.453 (0.018)	9.014 (0.097)	0.186 (0.007)
	CompLM	2.006 (0.171)	5.002 (0.061)	0.359 (0.271)	0.351 (0.246)	6.238 (0.177)	0.454 (0.019)	31.036 (1.34)	0.349 (0.017)

Table 2: Means and standard derivations (in brackets) of the estimated expansion coefficients for the functions and re-transformed coefficients for compositions with $(N, n_i) = (100, 60)$. The ideal values of $\widehat{\boldsymbol{\lambda}}_k = (\hat{\lambda}_{k1}, \hat{\lambda}_{k2}, \cdots, \hat{\lambda}_{k7})'$ $(k = 1, 2)$ are $\widehat{\boldsymbol{\lambda}}_1 = (3, 2, \cdots, -3)'$ and $\widehat{\boldsymbol{\lambda}}_2 = (-3, -2, \cdots, 3)'$, and those of $\widehat{\boldsymbol{\gamma}}_k$ are $\boldsymbol{\gamma}_k$ $(k = 1, 2)$.

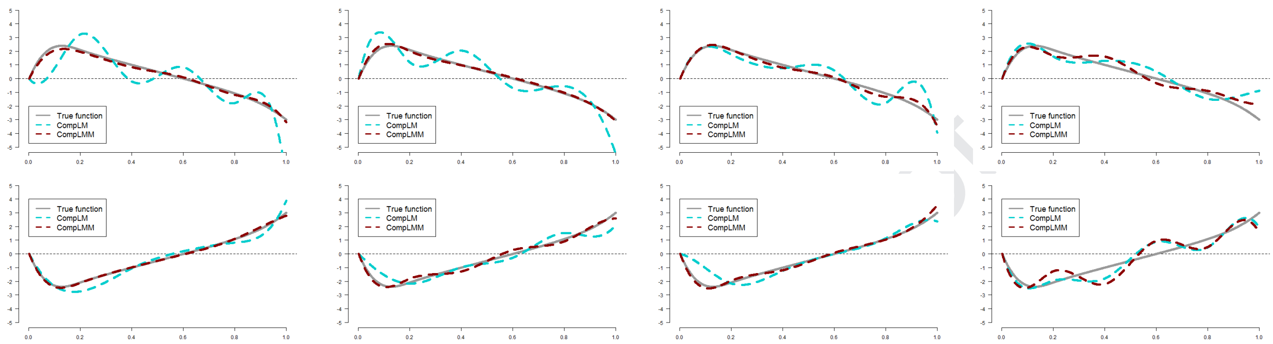
Model	Coefficients	Functional							Compositional		
	$\boldsymbol{\lambda}_k$ / $\boldsymbol{\gamma}_k$	$\hat{\lambda}_{k1}$	$\hat{\lambda}_{k2}$	$\hat{\lambda}_{k3}$	$\hat{\lambda}_{k4}$	$\hat{\lambda}_{k5}$	$\hat{\lambda}_{k6}$	$\hat{\lambda}_{k7}$	$\hat{\gamma}_{k1}$	$\hat{\gamma}_{k2}$	$\hat{\gamma}_{k3}$
					$\sigma = 0.5$						
CompLMM	$k = 1$	2.962 (0.485)	1.996 (0.509)	1.003 (0.401)	0.008 (0.419)	−1.026 (0.564)	−1.989 (0.599)	−3.004 (0.472)	0.786 (0.055)	0.101 (0.018)	0.112 (0.041)
	$k = 2$	−2.947 (0.377)	−2.047 (0.468)	−0.966 (0.4)	−0.031 (0.415)	1.05 (0.553)	1.936 (0.581)	3.039 (0.459)	0.2 (0.001)	0.2 (0.001)	0.6 (0.002)
CompLM	$k = 1$	2.946 (3.154)	2.063 (3.848)	0.991 (3.239)	−0.059 (3.412)	−0.841 (4.609)	−2.246 (5.008)	−2.719 (4.059)	0.785 (0.059)	0.103 (0.021)	0.113 (0.043)
	$k = 2$	−2.755 (3.141)	−2.297 (3.892)	−0.806 (3.312)	−0.181 (3.593)	1.272 (4.908)	1.964 (5.147)	2.78 (4.095)	0.2 (0.013)	0.199 (0.012)	0.6 (0.017)
					$\sigma = 1$						
CompLMM	$k = 1$	2.969 (0.804)	1.981 (0.931)	1.011 (0.753)	0.01 (0.806)	−1.029 (1.102)	−2.008 (1.19)	−2.987 (0.934)	0.786 (0.055)	0.102 (0.018)	0.112 (0.041)
	$k = 2$	−2.911 (0.767)	−2.072 (0.961)	−0.953 (0.816)	−0.044 (0.835)	1.09 (1.098)	1.868 (0.159)	3.086 (0.924)	0.2 (0.003)	0.2 (0.003)	0.6 (0.004)
CompLM	$k = 1$	2.927 (3.204)	2.078 (3.926)	0.981 (3.306)	−0.044 (3.473)	−0.864 (4.683)	−2.233 (5.087)	−2.725 (4.146)	0.785 (0.059)	0.103 (0.021)	0.113 (0.043)
	$k = 2$	−2.716 (3.17)	−2.339 (3.938)	−0.775 (3.371)	−0.206 (3.654)	1.306 (4.994)	1.92 (5.249)	2.812 (4.178)	0.2 (0.013)	0.199 (0.012)	0.6 (0.018)
					$\sigma = 1.5$						
CompLMM	$k = 1$	2.988 (1.161)	1.962 (1.379)	1.032 (1.129)	−0.004 (1.216)	−1 (1.673)	−2.058 (1.808)	−2.956 (1.411)	0.786 (0.055)	0.102 (0.018)	0.112 (0.041)
	$k = 2$	−2.889 (1.167)	−2.089 (1.456)	−0.951 (1.233)	−0.049 (1.255)	1.118 (1.644)	1.81 (1.725)	3.128 (1.386)	0.2 (0.005)	0.2 (0.004)	0.6 (0.006)
CompLM	$k = 1$	2.909 (3.288)	2.094 (4.043)	0.97 (3.406)	−0.03 (3.57)	−0.888 (4.807)	−2.22 (5.221)	−2.731 (4.275)	0.784 (0.059)	1.03 (0.021)	0.113 (0.043)
	$k = 2$	−2.676 (3.236)	−2.381 (4.03)	−0.745 (3.466)	−0.231 (3.749)	1.34 (5.126)	1.876 (5.399)	2.845 (4.299)	0.2 (0.013)	0.199 (0.013)	0.6 (0.018)
					$\sigma = 3$						
CompLMM	$k = 1$	3.022 (2.274)	1.952 (2.726)	1.04 (2.252)	0.009 (2.405)	−1.025 (3.346)	−2.076 (3.618)	−2.937 (2.796)	0.785 (0.056)	0.102 (0.019)	0.112 (0.041)
	$k = 2$	−2.803 (2.277)	−2.167 (2.842)	−0.933 (2.411)	−0.031 (2.466)	1.123 (3.25)	1.733 (3.412)	3.197 (2.774)	0.2 (0.009)	0.2 (0.008)	0.6 (0.012)
CompLM	$k = 1$	2.853 (3.716)	2.139 (4.6)	0.937 (3.873)	0.014 (4.053)	−0.96 (5.444)	−2.181 (5.907)	−2.748 (4.867)	0.784 (0.06)	0.103 (0.021)	0.113 (0.043)
	$k = 2$	−2.559 (3.637)	−2.507 (4.544)	−0.653 (3.943)	−0.306 (4.221)	1.443 (5.762)	1.743 (6.103)	2.944 (4.861)	0.2 (0.015)	0.2 (0.014)	0.6 (0.021)

no remarkable difference.

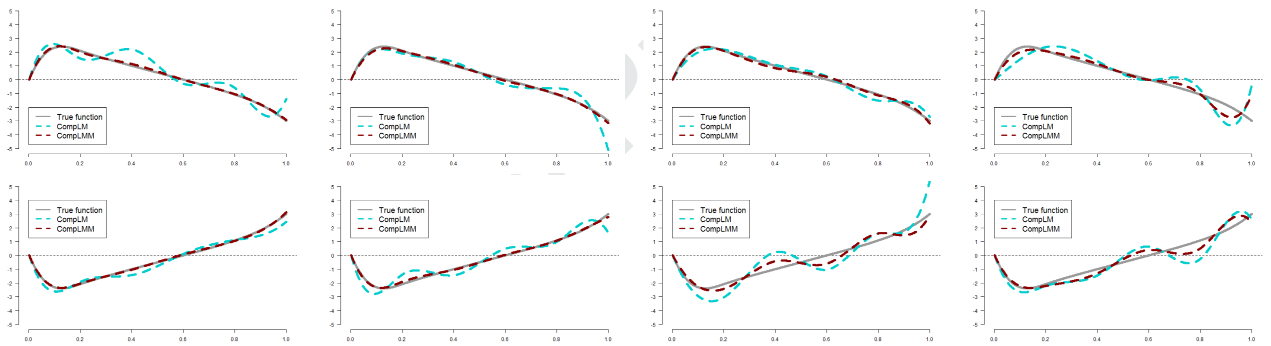
As exemplified in Fig. 1, the curves (in red) of the estimated functional coefficients from CompLMM move closer to the true settings (in gray) than those (in cyan) from CompLM. For both the increasing and the decreasing cases, CompLMM fits the functions well across the interval, whereas CompLM, despite capturing the general trends of the functions, creates relatively large periodic perturbations. Moreover, the biases between the true and fitted curves from the two models are eliminated gradually as the sample size increases (e.g., $(N, n_i) = (300, 90)$).



(a) $(N, n_i) = (100, 60)$.



(b) $(N, n_i) = (300, 60)$.



(c) $(N, n_i) = (300, 90)$.

Fig. 1: Curves of the estimated functional coefficients. The columns from left to right denote the four levels of the SNRs from $\sigma = 0.5$ to $\sigma = 3$. The upper and lower sub-rows indicate the two functional covariates.

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

5 Application

In this section, we adopt the proposed CompLMM in a real data study to demonstrate its usefulness. The existing approach to complex data modeling, namely CompLM (Wang et al., 2015, 2019a), is also used for comparison purposes.

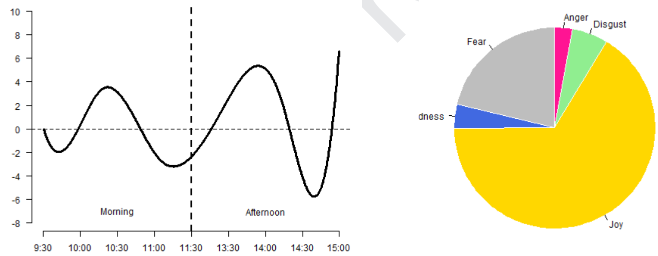
Using the case of China's stock market, we aim to measure the influence of indirect information on stock prices as well as the historical price trend. As exemplified by numerous studies, macroeconomic indicators (Chen et al., 1986), public online emotion (Ruan et al., 2018; Zhou et al., 2017), and analysts' recommendations (Duan et al., 2013) may improve the interpretability and accuracy of models for this problem. 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' emotions (OIE) 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. Stocks that have fewer than 40 active trading days are omitted, and finally 271 stocks are left. Specifically, both the DCP and the DV are scalar, and the IPR recorded every five minutes is described as a smoothing curve from opening to closing. Moreover, the data on the OIE are measured by Zhou et al. (2017), where observations are naturally of a compositional structure associated with five types of emotions labeled "Anger," "Disgust," "Joy," "Sadness," and "Fear." Thus, the regression contains two scalar covariates (including the intercept), one functional covariate, and one compositional covariate.

Then, we conduct CompLMM with all four covariates in the random effects as well as CompLM. Specifically, the IPR on each trading day is separately represented by seven expansion coefficients under the B-spline basis functions ϕ described in Section 4 over $[0, 1]$, where 0 and 1 indicate the opening (9:30 a.m.) and closing (15:00 p.m.) times, respectively. CompLM shows a poor result in this regression, as the variance of the residuals from it reaches an unacceptable level (460.43). By contrast, the introduction of the random effects in CompLMM reduces that variance to only 4.02, which makes for a reliable interpretation. Table 3 reports the estimated results for the fixed effects from the two models and Fig. 2 presents the related curves and pie charts of the estimated functional and compositional coefficients, respectively.

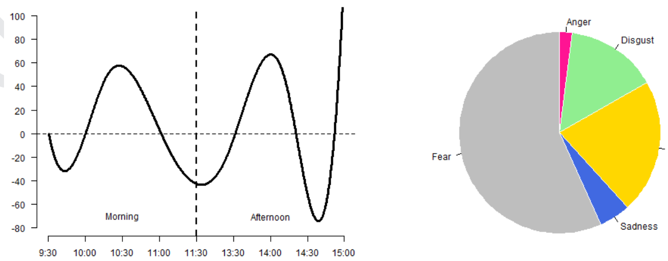
As shown in Table 3, the contribution of the DV to the stock price is contrasting in the two models: positive (0.09) in CompLMM and negative (-0.13) in CompLM. However, the absolute values of both coefficients are low, implying that the DV may not have a significant influence on the stock price. For the IPR, the directions of the estimated expansion coefficients from the two models are close in general,with six of the seven components having a consistent sign. As displayed in the left column of Fig. 2, the two images of the functional coefficients share a similar shape and the returns near 10:30 a.m., 14:00 p.m., and the closing time show relatively high marginal effects on the stock price. The difference between the two curves is that the range of values in CompLM is around 10 times larger than that in CompLMM, which accounts for the bad performance of CompLM to a great extent. Finally, as presented in the right column of Fig. 2, the two models differ in the estimated compositional coefficient for the OIE, although they both consider "Joy" and "Fear" to be two important emotions for explaining the DCP. In CompLMM, "Joy" has the largest influence on the stock price (proportion of 0.66), with "Fear" second (0.21); however, these two emotions change places in CompLM: 0.21 for "Joy" vs. 0.57 for "Fear". Since the increase in the inner part of a composition implies a general decrease in the others, it is hard to measure the influence of a specific part separately (Pawlowsky-Glahn et al., 2015). Hence, we only briefly discuss the marginal contribution of each type of emotion in the regression.

Table 3: Estimated coefficients for the fixed effects in the real data study. The estimated functional coefficient for the IPR is reported by its expansion coefficients, as indicated by the sub-columns ϕ_j $(j = 1, 2, \cdots, 7)$.

Model	Intercept	DV	IPR							Anger	Disgust	Joy	Sadness	Fear
			ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7					
CompLMM	15.81	0.09	-5.58	10.24	-7	1.13	11.11	-15.49	9.58	0.03	0.06	0.66	0.04	0.21
CompLM	21.29	-0.13	-84.56	130.84	-29.2	-74.13	199.59	-233.4	137.69	0.02	0.15	0.21	0.05	0.57



(a) CompLMM.



(b) CompLM.

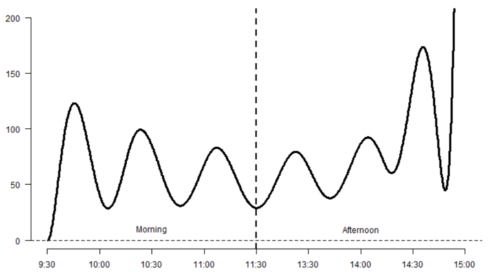
Fig. 2: Curves and pie charts of the estimated functional and compositional coefficients for the IPR and OIE, respectively. The vertical dotted line in the curve divides the trading day into morning and afternoon.

Next, we focus on the estimated results for the random effects for CompLMM, as reported in Table 4. The large variance of the intercept (314.3) indicates that the stocks involved have great differences in prices. The variance of the DV is relatively small (i.e., only 0.1), which implies that its influence has few changes across stocks; therefore, the DV can be regarded as an inessential factor in this case. To describe the overall variation of the functional coefficient, we plot the variance function based on the covariance matrix of the seven expansion coefficients, where the point-wise variances during all trading hours exceed 50 in general, especially those before 10:00 a.m. and after 14:30 p.m. This result verifies that past trends of the return from different stocks have different influences on their prices in the future. Finally, we sum the variances of the four ilr coordinates and obtain the related total variance of the compositional covariate for the OIE as 36.3. This result verifies that indirect information such as the OIE, although shared by all stocks, can also enhance the performance of the regression model for the stock price in various ways.

Table 4: Estimated covariance matrix of the random effects for CompLMM in the real data study. The sub-columns ϕ_j $(j = 1, 2, \cdots, 7)$ are the same as in Table 3 and w^*_{1k} $(k = 1, 2, \cdots, 4)$ indicate the ilr coordinates of the compositional coefficient for the OIE. The variances are highlighted in bold. The variance function and total covariance of the functional and compositional coefficients for the IPR and OIE are also plotted and reported.

	Intercept	DV	IPR							OIE			
			ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	w^*_{11}	w^*_{12}	w^*_{13}	w^*_{14}
Intercept	**314.6**	2.9	60.8	-73.6	20.8	38.4	-63.4	163.4	40.6	15.3	-1.3	-24.6	1.2
DV		**0.1**	0.5	-0.8	0.6	-0.2	-0.6	1.3	-1	0.7	0.2	-0.6	0.4
ϕ_1			**669.9**	-743.6	442.9	-282.7	245.5	-195.4	106.8	3.2	4.3	-8	1.2
ϕ_2				**1039.1**	-761.5	562.9	-533.9	418.3	-233.6	-6.5	-2.7	12.8	-4.5
ϕ_3					**693.1**	-608.3	599.1	-444.7	239.2	6.1	1.7	-9.6	8.1
ϕ_4						**680.1**	-776.4	586.6	-306.5	-3.7	-1.9	4.5	-10.5
ϕ_5							**1152.1**	-1151.5	727.9	-8.9	-0.4	13.6	-5.3
ϕ_6								**1637.4**	-1348.7	24.6	7.8	-31.3	30.7
ϕ_7									**1362.5**	-23.2	-11.8	25	-31.9
w^*_{11}										**7.1**	2.8	-8.8	7.5
w^*_{12}											**5.6**	-2.3	2.8
w^*_{13}												**13.2**	-10.1
w^*_{14}													**10.4**

Variance function for IPR



Total variance for OIE: 36.3

In conclusion, the real data study illustrates the potential of the proposed CompLMM for longitudinal complex data from an application perspective. Introducing 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 direct and indirect indicators for China's stock market. These issues need to be addressed in future work.

6 Discussion

This study investigates an LMM technique for longitudinal complex data named CompLMM, involving scalar continuous response and complex data covariates with diversified characteristics. Through 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 shows a significant improvement in fitting responses. Following the linear framework of complex data modeling, CompLMM first unifies the numeric representation of different types of variables such that the traditional LMM can then be conducted to obtain the intermediate results and transform them back to have related diversified features. This model also encourages a more comprehensive interpretation for regression on complex data. Moreover, some theoretical properties are also presented that support the computational procedure of the parameter estimation for CompLMM. As illustrated by both the numerical experiment and the real data study, the proposed CompLMM succeeds in dealing with longitudinal complex data and efficiently estimating the parameters with more reliable response fittings.

We focus on the parameter estimation and its general interpretation for the proposed CompLMM. However, the trade-off between the accuracy and interpretation of the proposed model also needs due consideration, to which many solutions for traditional scalar covariates have been proposed. These statistical methods provide instructive strategies for selecting random effects with diversified characteristics, which face great challenges in theory but deserve further research. Meanwhile, 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 related 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 for multiple types of complex data, with functional, compositional, and other more complicated features, are also an important and challenging issue in regression. Although empirical methods (e.g., the bootstrap) have offered partial solutions to this problem, related hypothesis tests for complex data such as the function with an infinite dimension and composition involving constraints, should also be developed.

Appendix: More results from the numerical experiment

See Tables 5 and 6.

Table 5: Means and standard derivations (in brackets) of the estimated expansion coefficients for the functions and re-transformed coefficients for the compositions with $(N, n_i) = (300, 60)$. The ideal values are the same as in Table 2.

Model	Coefficients	Functional							Compositional		
	β_k / γ_k	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	γ_{k1}	γ_{k2}	γ_{k3}
					$\sigma = 0.5$						
CompLMM	$k = 1$	2.988 (0.257)	2.011 (0.287)	0.983 (0.213)	0.004 (0.226)	−1.003 (0.308)	−2.007 (0.325)	−2.971 (0.25)	0.795 (0.033)	1.01 (0.012)	0.104 (0.023)
	$k = 2$	−3.009 (0.216)	−1.985 (0.269)	−1.015 (0.222)	0.02 (0.231)	0.982 (0.311)	2.018 (0.32)	2.973 (0.248)	0.2 (0.001)	0.2 (0.001)	0.6 (0.001)
CompLM	$k = 1$	3.257 (01.956)	1.675 (2.439)	1.241 (1.984)	−0.257 (2.019)	−0.674 (2.701)	−2.224 (2.942)	−2.861 (2.369)	0.796 (0.034)	0.101 (0.013)	0.103 (0.024)
	$k = 2$	−3.039 (1.751)	−1.895 (2.136)	−1.121 (1.76)	0.172 (1.836)	0.774 (2.506)	2.195 (2.609)	2.885 (2.132)	0.199 (0.007)	0.2 (0.007)	0.6 (0.01)
					$\sigma = 1$						
CompLMM	$k = 1$	2.972 (0.424)	2.024 (0.526)	0.974 (0.417)	0.01 (0.447)	−1.007 (0.608)	−2.019 (0.637)	−2.95 (0.49)	0.795 (0.033)	0.101 (0.012)	0.2 (0.002)
	$k = 2$	−3.02 (0.427)	−1.973 (0.53)	−1.028 (0.436)	0.036 (0.457)	0.969 (0.615)	2.032 (0.634)	2.955 (0.494)	0.2 (0.002)	0.2 (0.001)	0.6 (0.002)
CompLM	$k = 1$	3.24 (2.004)	1.688 (2.499)	1.232 (2.03)	−0.251 (2.065)	−0.679 (2.762)	−2.233 (3.001)	−2.843 (2.421)	0.796 (0.034)	0.101 (0.013)	0.103 (0.024)
	$k = 2$	−3.044 (1.776)	−1.887 (2.17)	−1.131 (1.787)	0.186 (1.866)	0.763 (2.562)	2.209 (2.664)	2.868 (2.161)	0.199 (0.007)	0.2 (0.007)	0.6 (0.01)
					$\sigma = 1.5$						
CompLMM	$k = 1$	2.956 (0.609)	3.224 (2.068)	1.701 (2.581)	1.224 (2.094)	−0.245 (2.133)	−0.684 (2.852)	−2.241 (3.09)	0.796 (0.034)	0.101 (0.013)	0.104 (0.023)
	$k = 2$	−3.029 (0.637)	−1.962 (0.792)	−1.041 (0.652)	0.055 (0.684)	0.951 (0.923)	2.05 (0.95)	2.935 (0.741)	0.2 (0.002)	0.2 (0.002)	0.6 (0.003)
CompLM	$k = 1$	2.956 (0.609)	2.036 (0.771)	0.966 (0.623)	0.017 (0.667)	−1.011 (0.907)	−2.031 (0.948)	−2.93 (0.726)	0.795 (0.033)	0.101 (0.012)	0.104 (0.023)
	$k = 2$	−3.05 (1.825)	−1.88 (2.233)	−1.141 (1.838)	0.201 (1.922)	0.751 (2.65)	2.223 (2.753)	2.851 (2.216)	0.199 (0.007)	0.2 (0.007)	0.6 (0.011)
					$\sigma = 3$						
CompLMM	$k = 1$	2.918 (1.183)	2.064 (1.515)	0.945 (1.241)	0.037 (1.317)	−1.025 (1.789)	−2.065 (1.869)	−2.864 (1.435)	0.795 (0.034)	0.101 (0.012)	0.104 (0.023)
	$k = 2$	−3.061 (1.261)	−1.924 (1.572)	−1.084 (1.297)	0.116 (1.364)	0.889 (1.836)	2.117 (1.879)	2.865 (1.458)	0.2 (0.005)	0.2 (0.004)	0.6 (0.007)
CompLM	$k = 1$	3.176 (2.348)	1.739 (2.938)	1.198 (2.383)	−0.227 (2.446)	−0.698 (3.276)	−2.266 (3.513)	−2.774 (2.822)	0.796 (0.035)	0.101 (0.013)	0.103 (0.024)
	$k = 2$	−3.067 (2.088)	−1.859 (2.572)	−1.172 (2.115)	0.244 (2.217)	0.717 (3.08)	2.264 (3.188)	2.8 (2.515)	0.199 (0.009)	0.2 (0.008)	0.6 (0.012)

Table 6: Means and standard derivations (in brackets) of the estimated expansion coefficients for the functions and re-transformed coefficients for the compositions with $(N, n_i) = (300, 90)$. The ideal values are the same as in Table 2.

Model	Coefficients	Functional							Compositional		
	β_k / γ_k	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	γ_{k1}	γ_{k2}	γ_{k3}
					$\sigma = 0.5$						
CompLMM	$k = 1$	3.017 (0.246)	1.985 (0.224)	1.021 (0.179)	−0.021 (0.187)	−0.989 (0.246)	−2.003 (0.248)	−2.999 (0.198)	0.798 (0.035)	0.1 (0.011)	0.102 (0.026)
	$k = 2$	−2.974 (0.163)	−2.031 (0.2)	−0.976 (0.173)	−0.021 (0.188)	1.019 (0.256)	1.991 (0.269)	2.998 (0.209)	0.2 (0.001)	0.2 (0.001)	0.6 (0.001)
CompLM	$k = 1$	3.132 (1.58)	1.834 (1.964)	1.195 (1.675)	−0.211 (1.726)	−0.818 (2.298)	−2.1 (2.364)	−2.987 (1.86)	0.798 (0.036)	0.1 (0.012)	0.102 (0.026)
	$k = 2$	−3.018 (1.497)	−1.916 (1.752)	−1.093 (1.432)	0.077 (1.541)	0.893 (2.112)	2.137 (2.288)	3.022 (1.781)	0.199 (0.006)	0.2 (0.006)	0.6 (0.009)
					$\sigma = 1$						
CompLMM	$k = 1$	3.03 (0.372)	1.967 (0.406)	1.043 (0.339)	−0.042 (0.361)	−0.975 (0.481)	−2.005 (0.484)	−3.002 (0.389)	0.797 (0.035)	0.1 (0.011)	0.102 (0.026)
	$k = 2$	−2.953 (0.324)	−2.062 (0.398)	−0.952 (0.345)	−0.043 (0.374)	1.039 (0.508)	1.98 (0.535)	3.001 (0.417)	0.2 (0.001)	0.2 (0.01)	0.6 (0.001)
CompLM	$k = 1$	3.145 (1.594)	1.817 (1.976)	1.216 (1.682)	−0.231 (1.734)	−0.804 (2.313)	−2.102 (2.377)	−2.989 (1.873)	0.798 (0.036)	0.1 (0.012)	0.102 (0.026)
	$k = 2$	−3.001 (1.518)	−1.942 (1.789)	−1.072 (1.474)	0.058 (1.586)	0.909 (2.16)	2.131 (2.329)	3.019 (1.817)	0.2 (0.006)	0.2 (0.006)	0.6 (0.009)
					$\sigma = 1.5$						
CompLMM	$k = 1$	3.043 (0.513)	1.948 (0.597)	1.067 (0.501)	−0.064 (0.537)	−0.959 (0.717)	−2.009 (0.721)	−3.004 (0.581)	0.797 (0.035)	0.1 (0.011)	0.102 (0.026)
	$k = 2$	−2.931 (0.485)	−2.092 (0.596)	−0.929 (0.517)	−0.065 (0.56)	1.059 (0.759)	1.971 (0.799)	3.004 (0.624)	0.2 (0.002)	0.2 (0.002)	0.6 (0.003)
CompLM	$k = 1$	3.158 (1.621)	1.801 (2.007)	1.237 (1.704)	−0.25 (1.76)	−0.791 (2.35)	−2.104 (2.411)	−2.991 (1.904)	0.798 (0.036)	0.1 (0.012)	0.102 (0.026)
	$k = 2$	−2.984 (1.554)	−1.967 (1.846)	−1.051 (1.532)	0.04 (1.651)	0.926 (2.232)	2.124 (2.396)	3.016 (1.873)	0.2 (0.006)	0.2 (0.006)	0.6 (0.009)
					$\sigma = 3$						
CompLMM	$k = 1$	3.086 (0.953)	1.885 (1.175)	1.142 (0.988)	−0.135 (1.062)	−0.906 (1.419)	−2.023 (1.423)	−3.009 (1.153)	0.797 (0.035)	0.1 (0.011)	0.102 (.026)
	$k = 2$	−2.872 (0.967)	−2.176 (1.191)	−0.863 (1.036)	−0.125 (1.117)	1.113 (1.511)	1.948 (1.586)	3.004 (1.244)	0.2 (0.004)	0.2 (0.004)	0.6 (0.006)
CompLM	$k = 1$	3.197 (1.778)	1.751 (2.198)	1.299 (1.855)	−0.309 (1.931)	−0.752 (2.586)	−2.109 (2.636)	−2.996 (2.099)	0.797 (0.036)	0.1 (0.012)	0.102 (0.026)
	$k = 2$	−2.932 (1.744)	−2.045 (2.112)	−0.989 (1.791)	−0.016 (1.931)	0.975 (2.577)	2.104 (2.731)	3.007 (2.143)	0.2 (0.004)	0.2 (0.004)	0.6 (0.006)

References

Aitchison J., 1982. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Statistical Methodological) 44, 139–177.

Aitchison J., 1986. The statistical analysis of compositional data. Chapman and Hall London.

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, 1–19.

Bruno F., Greco F., Ventrucci M., 2016. Non-parametric regression on compositional covariates using bayesian P-splines. Statistical Methods & 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.

Di Marzio M., Panzera A., Venieri C., 2015. Non-parametric regression for compositional data. Statistical Modelling 15, 113–133.

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., Pawlowsky-Glahn V., Mateu-Figueras G., Barcelo-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. The Annals of Statistics 43, 2296–2325.

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 (Statistical Methodology) 78, 637–653.

Hall P., Horowitz J.L., 2007. Methodology and convergence rates for functional linear regression. The 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., Ware J.H., 1982. Random-effects models for longitudinal data. Biometrics 38, 963–974.

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.

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. John Wiley & Sons.

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 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 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 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. The Canadian Journal of Statistics 37, 266–281.

Zhou Z., Ke X., Zhao J., 2017. Tales of emotion and stock in China: Volatility, causality and prediction. World Wide Web 21, 1–24.