*Article*

# Optimal Estimation of Large Functional and Longitudinal Data by Using Functional Linear Mixed Model

**Mengfei Ran** [1] and **Yihe Yang** [2,*]

1 Graduate School of Engineering Science, Osaka University, Osaka 560-0043, Japan
2 Department of Population and Quantitative Health Science, Case Western Reserve University, Cleveland, OH 44106, USA
* Correspondence: yxy1234@case.edu

**Abstract:** The estimation of large functional and longitudinal data, which refers to the estimation of mean function, estimation of covariance function, and prediction of individual trajectory, is one of the most challenging problems in the field of high-dimensional statistics. Functional Principal Components Analysis (FPCA) and Functional Linear Mixed Model (FLMM) are two major statistical tools used to address the estimation of large functional and longitudinal data; however, the former suffers from a dramatically increasing computational burden while the latter does not have clear asymptotic properties. In this paper, we propose a computationally effective estimator of large functional and longitudinal data within the framework of FLMM, in which all the parameters can be automatically estimated. Under certain regularity assumptions, we prove that the mean function estimation and individual trajectory prediction reach the minimax lower bounds of all nonparametric estimations. Through numerous simulations and real data analysis, we show that our new estimator outperforms the traditional FPCA in terms of mean function estimation, individual trajectory prediction, variance estimation, covariance function estimation, and computational effectiveness.

**Keywords:** functional data analysis; functional linear mixed model; functional principal components analysis; longitudinal data analysis

**MSC:** 62G20; 62J10; 62R10

## 1. Introduction

Functional Data Analysis (FDA) is a field of statistics that examines data to provide information about curves, surfaces, or anything else that varies along a continuum [1]. The concrete continuum of these functions can be time, spatial location, wavelength, and probability, with applications to biomechanics, biomedicine, ecology, epidemiology, and neurology, among many others [2,3]. Functional data can be classified into common design data and independent design data, where the sampling locations of all individuals are the same under the former setting, while the sampling locations are individual-wise under the latter setting [4]. Longitudinal data can be regarded as a specific type of functional data where the underlying continuum is just time [5]. In the field of longitudinal data analysis, the counterparts of common design and independent design are called balanced longitudinal data and unbalanced longitudinal data, respectively.

The coronavirus 2019 (COVID-19) data is typical balanced longitudinal data, which records the daily infections (transformed in logarithm) of the COVID-19 pandemic from 51 states of the United States (US), where the date is from 16 March 2020 to 14 August 2020 (Johns Hopkins COVID-19 Case Tracker: https://www.kaggle.com/datasets/thecansin/johns-hopkins-covid19-case-tracker?resource=download). Figure 1 demonstrates the profiles of daily increasing infections and the related trajectories predictions separately yielded with the R package mgcv [6] of six states in the US. From this figure, it is easy to see some

basic features of large longitudinal data: (1) functionality, the daily infections of states are likely some smooth curves of time; (2) heterogeneity, different individuals usually have different trends over time; and (3) high-dimensionality, the number of observations ($T = 152$) much larger than the number of individuals ($N = 51$). How to deal with this kind of large functional and longitudinal data is one of the most challenging problems in the field of high-dimensional statistics.
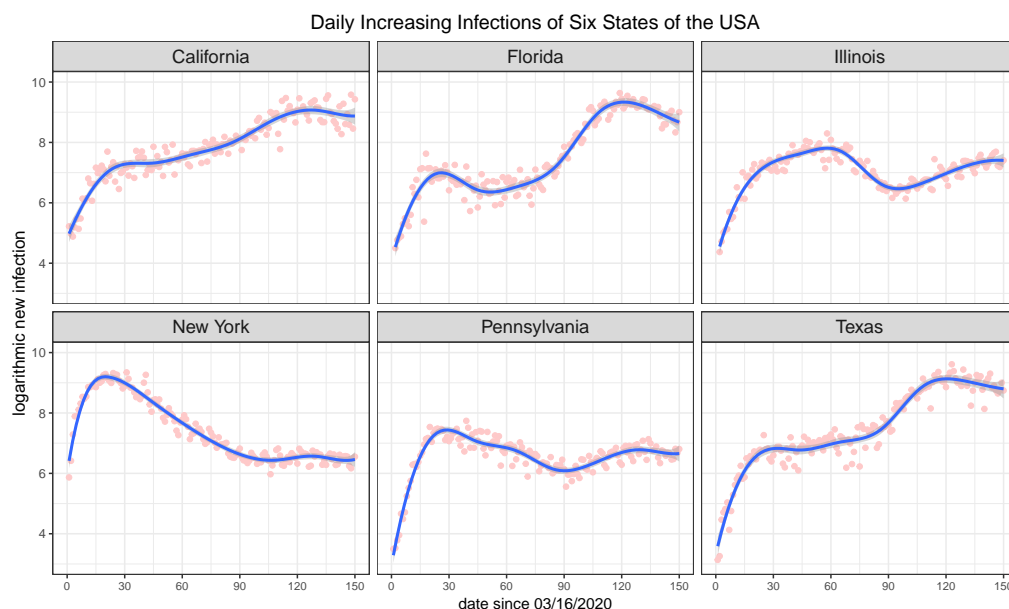


**Figure 1.** Illustration of the COVID-19 data. The X-axis records the dates from 16 March 2020 to 14 August 2020 and the Y-axis shows the numbers of the logarithm of the daily infections.

Functional Principal Components Analysis (FPCA) and Functional Linear Mixed Model (FLMM) are two common approaches that estimate the dominant modes of variation of a sample of random trajectories around an overall mean function. FPCA and FLMM share the same principle to estimate the mean function, but apply totally different strategies to estimate the covariance function and predict the individual trajectory. Specifically, FPCA first estimates the two-dimensional covariance function, then yields the eigenfunctions through PCA, and finally predicts the individual trajectory with the estimated eigenfunctions. In contrast, FLMM parameterizes the individual trajectories through basis functions, in particular, it specifies a Gaussian prior distribution on the coefficients of basis functions, then estimates the covariance matrix of these random coefficients using REstricted Maximum Likelihood (REML) [7], and eventually predicts the individual trajectories through the Best Linear Unbiased Prediction (BLUP) [8] and yields the covariance function of random curves using a weighted sum of inner products of basis functions. Commonly-employed nonparametric techniques for FPCA include the kernel method and local polynomial modeling ([9–11] and smoothing splines ([12–16]), while frequently-used basis expansion functions include B-splines ([17–20]), wavelets ([21–23]), a combination of linear mixed-effects modeling and local polynomial smoothing ([24,25]), etc. See more aspects of functional data in the following monographs ([1,26]) and review papers ([3,27,28]).

One of the greatest advantages of FPCA may be its clear asymptotic properties. Under mild regularity conditions, Yao et al. [11] showed the convergence rates of estimates of mean function covariance function, variance, eigenfunctions, and eigenvalues. Li and Hsing [10] further derived the strong uniform convergence rates of these estimates. Cai and Yuan [4] investigated the minimax risks of the mean function and covariance function estimates and Cai and Yuan [29] showed the minimax risk of mean function estimate in functional linear regression. However, because FPCA needs to estimate a two-dimensional

covariance surface, its computational burden will dramatically increase as the number of observations $T$ rises. In addition, since most FPCA methods employ local polynomial modeling [30] as the estimator, it is very difficult to select the optimal bandwidth in the two-dimensional kernel function in practice. On the other hand, FLMM suffers from more problems compared with FPCA, preventing its practical application in the field of FDA. First, FLMM requires estimating the covariance matrix of the basis expansion coefficients (which are treated as normal variables in LMM), but the estimation of this covariance matrix is usually unreliable if the number of employed bases is large. James et al. [17] provided a dimension reduction to estimate this covariance matrix, which resulted in a more complex estimator. Besides, the asymptotic properties of FLMM are difficult to study. Proper regularity conditions and proof techniques are lacking for FLMM.

In this paper, we propose a computationally effective estimator of large functional and longitudinal data by using FLMM, in which all the parameters can be automatically estimated. We also investigate the large sample property of the proposed automatic and flexible estimation of large functional data. Under mild regularity conditions, we prove that the estimation error of the mean curve estimator is bounded by $O(N^{-1} + T^{-2s})$ and the prediction error of the individual trajectory estimator is bounded by $O(T^{-2s/(2s+1)})$, where $s > 1$ is a constant governing the smoothness of tge mean function and individual trajectory. In particular, both these two convergence rates reach the minimax lower boundaries of the mean function estimation and individual trajectory prediction derived by Cai and Yuan [4], meaning that our estimation enjoys the minimax efficiency. To the best of our knowledge, our work is the first to investigate the large sample property of FLMM estimation when both $T$ and $N$ are diverging.

The rest of the paper is organized as follows. In Section 2, we present the primal representation of FLMM to analyze the functional data. In Section 3, the new estimator of large balanced longitudinal data and its large sample property are illustrated. In Section 4, we exhibit the simulation studies. In Section 5, we apply our method to study the COVID-19 data. The concluding remarks are summarized in Section 6 and all the technical proofs of the theorems are relegated to the Appendix.

## 2. Preliminary

In this section, we introduce the functional data and the top two statistical methods, i.e., FPCA and FLMM, to deal with the functional data briefly.

### 2.1. Settings

For a vector $\boldsymbol{a} = (a_j)_{p \times 1}$, $||\boldsymbol{a}||_q = (\sum_{j=1}^{p} |a_j|^q)^{1/q}$ with $q \in [0, \infty]$. For a symmetric matrix $\mathbf{A} = (A_{ij})_{p \times p}$, $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximum and minimum eigenvalues of matrix $\mathbf{A}$, and $||\mathbf{A}||_q = \max\{||\mathbf{A}\boldsymbol{a}||_q, ||\boldsymbol{a}||_q = 1\}$. Besides, $a_n \asymp b_n$ if there are positive constants $c$ and $C$ such that $c \le a_n/b_n \le C$. In addition, $O(\cdot)$ and $o(\cdot)$ are the infinitely large and small quantities, respectively, while $O_P(\cdot)$ and $o_P(\cdot)$ mean that such relationships hold with a probability tending to 1.

### 2.2. Functional Data Model

Let $X(t)$ be a second-order stochastic process defined in the compact interval $\mathcal{T}$, which is usually set as $\mathcal{T} = [0, 1]$ for convenience. The covariance function of $X(t)$ is given by

$$C(s, t) = \mathrm{E}[\{X(s) - \mathrm{E}(X(s))\}\{X(t) - \mathrm{E}(X(t))\}], \quad (s, t) \in \mathcal{T} \times \mathcal{T}. \tag{1}$$

Besides, let $X_1(t), \cdots, X_N(t)$ be $N$ independent realizations of $X(t)$, $\{t_1, \ldots, t_T\}$ be $T$ sampling locations of the realizations, and the observations of $X_i(t_j)$ be

$$Y_{ij} = X_i(t_j) + \varepsilon_{ij}, \quad (i, j) \in \{1, \ldots, N\} \times \{1, \ldots, T\}, \tag{2}$$

where $\varepsilon_{ij}$ are independent and identically distributed (IID) random errors with mean 0 and variance $\sigma^2$. In addition, the sampling location is considered as fixed numbers in $\mathcal{T}$,

and $\{X_i(t_j)\}$ and $\{e_{ij}\}$ are two mutually independent groups. In a more general case, the sampling locations can vary from individual, i.e.,

$$Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij}, \quad (i,j) \in \{1,\ldots,N\} \times \{1,\ldots,T_i\}, \tag{3}$$

where $t_{ij}$ is a random sampling location and $T_i$ is the individual-wise sampling locations. In the literature, the setting of functional data subject to (2) is called common design, while the setting subject to (3) is termed independent design. In longitudinal data analysis where $t$ is just time, the data subject to (2) is named as balanced longitudinal data or panel data, while the data subject to (3) is named as the unbalanced longitudinal data.

### 2.3. Methodologies of FPCA and FLMM

FPCA is almost the most frequently used technique to deal with the functional data in which the Karhunen–Loève representation of the stochastic process plays a central role. Specifically, let covariance function $C(s,t)$ satisfy the definition of a Mercer kernel, i.e.,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} C(s_i, t_j)c_i c_j \geq 0 \tag{4}$$

for all finite sequence of points $(s_i, t_j) \in \mathcal{T} \times \mathcal{T}$ and all choices of real numbers $c_1, \ldots, c_n$. Then, according to Mercer's theorem, there consequently exists a series of eigenfunctions $\{\phi_k(t)\}$ and a series of non-increasing eigenvalues $\{\lambda_k\}$ satisfying $\sum_{k=1}^{\infty} \lambda_k < \infty$ such that

$$C(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(t)\phi_k(s), \quad (s,t) \in \mathcal{T} \times \mathcal{T}. \tag{5}$$

In particular, under the condition of Mercer kernel, the Karhunen–Loève representation theory is able to guarantee that the $i$th realization of $X(t)$ can be expressed as

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t), \quad t \in \mathcal{T}, \tag{6}$$

where $\mu(t)$ is termed the mean function and $\xi_{i1}, \ldots, \xi_{ik}, \ldots$ are uncorrelated random coordinates with mean 0 and variance $E(\xi_{ik}^2) = \lambda_k$.

FLMM deals with the functional data in a more direct way than FPCA. Specifically, FLMM expresses each random function by some known basis functions:

$$X_i(t) = \sum_{k=1}^{p} \Psi_k(t)b_{ik} + \delta_i(t), \tag{7}$$

where $\{\Psi_k(t)\}$ are a set of basis functions, $\{b_{ik}\}$ are the corresponding random coordinates, and $\delta_i(t)$ is the approximation bias. In practice, the basis functions can be the Tikhonov basis (as defined in $\mathcal{T} = [0,1]$):

$$\Psi_k(t) = \sqrt{2}\cos((k-1)\pi t). \tag{8}$$

Alternatively, B-splines [31] are common bases that are generated in an iterative procedure. If $\delta_i(t)$ is negligible, the uncertainty of random function $X_i(t)$ is almost described by the random coordinates $\{b_{ik}\}$. In the literature, it is routine to assume

$$\begin{pmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{ip} \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \cdots & \Gamma_{1p} \\ \Gamma_{21} & \Gamma_{22} & \cdots & \Gamma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{p1} & \Gamma_{p2} & \cdots & \Gamma_{pp} \end{pmatrix} \right), \tag{9}$$

where $\beta_1, \ldots, \beta_p$ are the mean parameters and $\boldsymbol{\Gamma} = (\Gamma_{ij})_{p \times p}$ is a covariance matrix. As a result,

$$\mu(t) \approx \sum_{k=1}^{p} \Psi_k(t)\beta_k, \quad C(s,t) \approx \sum_{k=1}^{p}\sum_{h=1}^{p} \Gamma_{kh}\Psi_k(s)\Psi_h(t). \tag{10}$$

In addition, if $\{\Psi_k(t)\}$ are chosen as the eigenfunctions of the covariance function $C(s,t)$, the covariance matrix $\boldsymbol{\Gamma}$ will reduce into a diagonal matrix with the eigenvalues $\{\lambda_k\}$. On the other hand, FPCA can reduce to an FLMM if the eigenfunctions are rotated for a better practical interpretation, which is subject to a similar idea of factor analysis [32]. Generally, since $C(s,t)$ is never known in advance, FLMM will employ certain commonly-used basis functions and estimate $\boldsymbol{\Gamma}$ empirically.

### *2.4. Comparison between FPCA and FLMM*

FLMM is more efficient than FPCA in terms of both estimation and prediction. Due to the space limitation, we directly give the conclusions and show the estimation procedures of FPCA and FLMM in Appendix B.

We illustrate the features of FPCA from two aspects: estimation and prediction. For estimation, FLMM does not require estimating a two-dimensional covariance surface $C(s,t)$, which is the most time-consuming component for FPCA. Besides, all terms in FLMM can be automatically estimated using many mature and effective algorithms, such as the Newton–Raphson algorithm. Whereas the main approaches for FPCA to estimate the mean $\mu(t)$ and covariance $C(s,t)$ are local linear modeling [10], which is based on the Kernel function, it requires determining three different bandwidths $h_u, h_C, h_v$ (see details in Appendix B), which is supposedly very complex in practice. For prediction, FPCA will use all the observations in addition to the local linear coefficients $\hat{a}_0, \ldots, \hat{c}_1$, which are generated in the estimation process, to predict the values of $\mu(t)$, $X_{i*}(t)$, $C(s,t)$ at any new locations $t_{i*j}$ (and $s_{i*j}$), which is extremely costly in implementation. In contrast, predicting these functions is trivial for FLMM: it only needs to generate a new vector $\boldsymbol{\Psi}_{i*} = (\Psi(t_{i*1}), \ldots, \Psi(t_{i*T_{i*}}))^{\top}$ and then predict the values by using $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{b}}_i$, $\hat{\boldsymbol{\Gamma}}$ and $\hat{\sigma}^2$.

However, the most serious problem of FLMM is the lack of statistical theory in comparison with the numerous theoretical investigations of FPCA. One theoretical burden of FLMM is inherited from the LMM and generalized linear model (GLM, [33]): the statistical properties of LMM and GLM are different to analyze. As far as we are concerned, the tool to investigate the asymptotic properties of GLM is first given by Vonesh et al. [34], who pointed out that the convergence rate of the mean parameter is of $O(T_{\min}^{-1/2})$, where $T_{\min}$ is the minimum number of observations among all individuals. However, this convergence rate is not ideal because $T_{\min}$ can be less than 10 in many longitudinal data. Many works aim to improve the GLM estimation by removing the high-order bias of Laplace approximation, such as [35–37]; however, these bias-corrections make the asymptotic properties more difficult to analyze.

Another theoretical burden of FLMM is inherited from the basis expansion method, approximating a smooth function through a series of basis functions. As Ruppert et al. [38] once commented: "The literature on inference in smoothing is large and varied. Much of it is in the local polynomial or kernel smoothing context, where theoretical properties are more tractable. Inference for spline-based smoothing is less studied." Indeed, it is unclear under what conditions the FLMM using the basis expansion method is consistent. In the next section, we improve the traditional FLMM in terms of these two burdens.

## 3. Optimal Estimation for Large Balanced Longitudinal Data

In this paper, we propose a novel FLMM estimation for large balanced longitudinal data, which is effective in terms of computation and optimal in terms of statistical theory. Our new FLMM estimation for large balanced longitudinal data is motivated by the COVID-19 data, in which the number of sampling locations (i.e., the dates from 16 March 2020 to 14 August 2020) is much larger than the number of individuals (i.e., the 50 states plus

the District of Columbia). The traditional estimation of FLMM may encounter numerical instability and becomes extremely time-consuming because the complexity of the REML estimator dramatically increases with the dimension of the covariance matrix $\boldsymbol{\Sigma}_y$. In contrast, our new estimator for the variance components remains efficient and stable even when the dimension of $\boldsymbol{\Sigma}_y$ is very high. Note that under the setting of balanced longitudinal data, the sampling locations are the same for all individuals, i.e., $\boldsymbol{t} = (t_1, \ldots, t_T)^\top$, the covariance matrix $\boldsymbol{\Sigma}_{y_i}$ will become a universal one $\boldsymbol{\Sigma}_y$ and the basis function matrix $\boldsymbol{\Psi}_i$ will reduce to a universal matrix $\boldsymbol{\Psi}$ for all individuals.

### 3.1. Novel Estimating Procedure

The biggest difference between our new estimation and the traditional one is that we will use different numbers of basis functions to estimate the mean function and the individual trajectory. In order to guarantee both the mean function estimation and individual trajectory prediction meet the optimal convergence rates, we approximate the mean function and individual trajectory as follows. Specifically:

$$\mu(t) \approx \sum_{k=1}^{q} \Psi_k(t)\beta_k, \quad X_i(t) \approx \mu(t) + \sum_{k=1}^{p} \Psi_k(t)c_{ik}, \tag{11}$$

where $1 \le p \le q \le T$. Regarding the matrix form, we have

$$\boldsymbol{\mu} \approx \boldsymbol{\Psi}_q \boldsymbol{\beta}, \quad \boldsymbol{X}_i \approx \boldsymbol{\mu} + \boldsymbol{\Psi}_p \boldsymbol{c}_i, \tag{12}$$

where $\boldsymbol{\Psi}_q = (\Psi_k(t_j))_{T \times q}$, $\boldsymbol{\Psi}_p = (\Psi_k(t_j))_{T \times p}$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^\top$, $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{ip})^\top$. Here, the vector of random coordinates $\boldsymbol{c}_i$ is considered to follow a normal distribution

$$\boldsymbol{c}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma}), \tag{13}$$

where $\boldsymbol{\Gamma} = (\Gamma_{jk})$ is a $(p \times p)$ positive definite symmetric matrix. It is used to describe the local variability of individual trajectory $X_i(t)$ beyond the mean function $\mu(t)$.

The first step of our estimator is estimating the mean coefficient $\boldsymbol{\beta}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (\boldsymbol{Y}_i - \boldsymbol{\Psi}_q \boldsymbol{\beta})^\top \hat{\boldsymbol{\Sigma}}_y^{-1} (\boldsymbol{Y}_i - \boldsymbol{\Psi}_q \boldsymbol{\beta}) + \gamma \boldsymbol{\beta}^\top \mathbf{Q} \boldsymbol{\beta} \right\} \\
&= (N \boldsymbol{\Psi}_q^\top \hat{\boldsymbol{\Sigma}}_y^{-1} \boldsymbol{\Psi}_q + \gamma \mathbf{Q})^{-1} \boldsymbol{\Psi}_q^\top \hat{\boldsymbol{\Sigma}}_y^{-1} \left( \sum_{i=1}^{N} \boldsymbol{Y}_i \right),
\end{aligned} \tag{14}$$

where $\hat{\boldsymbol{\Sigma}}_y = \hat{\sigma}^2 \mathbf{I}_T + \boldsymbol{\Psi}_p \hat{\boldsymbol{\Gamma}} \boldsymbol{\Psi}_p^\top$, $\gamma$ is a smoothing parameter, and $\mathbf{Q}$ is a known semi-positive definite symmetric matrix ensuring

$$\int_{t \in \mathcal{T}} [\mu''(t)]^2 dt \approx \boldsymbol{\beta}^\top \mathbf{Q} \boldsymbol{\beta}. \tag{15}$$

Compared with traditional FLMM, which estimates $\mu(t)$ by (A51), we add a ridge penalty $\boldsymbol{\beta}^\top \mathbf{Q} \boldsymbol{\beta}$ to control the smoothness of $\mu(t)$. If $\Psi(t)$ is chosen as the Tikhonov bases, then $\mathbf{Q}$ is a diagonal matrix with the $j$th entry $P_{jj} = 2((j-1)\pi)^4$. If $\Psi(t)$ is chosen as the Bsplines, then $\mathbf{Q} = \mathbf{D}_2^\top \mathbf{D}_2$ where the expression of $\mathbf{D}_2$ can be found in Eilers and Marx [31]. In addition, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}) = (N \boldsymbol{\Psi}_q^\top \hat{\boldsymbol{\Sigma}}_y^{-1} \boldsymbol{\Psi}_q + \gamma \mathbf{Q})^{-1}. \tag{16}$$

When $T$ is large, we recommend the Sherman–Morrison–Woodbury formula to yield $\hat{\boldsymbol{\Sigma}}_y^{-1}$:

$$\hat{\boldsymbol{\Sigma}}_y^{-1} = \frac{1}{\hat{\sigma}^2} \left( \mathbf{I}_T - \boldsymbol{\Psi}_p (\boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p + \hat{\sigma}^2 \hat{\boldsymbol{\Gamma}}^{-1})^{-1} \boldsymbol{\Psi}_p^\top \right). \tag{17}$$

The mean function is then given by $\hat{\mu} = \mathbf{\Psi}_q \hat{\boldsymbol{\beta}}$ and its asymptotic covariance matrix is

$$\text{cov}(\hat{\mu}) = \mathbf{\Psi}_q (N\mathbf{\Psi}_q^\top \hat{\mathbf{\Sigma}}_y^{-1} \mathbf{\Psi}_q + \gamma \mathbf{Q})^{-1} \mathbf{\Psi}_q^\top. \tag{18}$$

The confidence band of $\hat{\mu}$ is constructed by using the diagonal elements of $\text{cov}(\hat{\mu})$.

The second step is to predict $c_i$ by using BLUP:

$$\begin{aligned}
\hat{c}_i &= \arg\min_{c_i} \left\{ \frac{1}{\hat{\sigma}^2} ||Y_i - \hat{\mu} - \mathbf{\Psi}_p c_i||_2^2 + c_i^\top \hat{\mathbf{\Gamma}}^{-1} c_i \right\} \\
&= (\mathbf{\Psi}_p^\top \mathbf{\Psi}_p + \hat{\sigma}^2 \hat{\mathbf{\Gamma}}^{-1}) \mathbf{\Psi}_p^\top (Y_i - \hat{u}).
\end{aligned} \tag{19}$$

This estimator is called BLUP because $\hat{c}_i$ can be regarded as the mean of the posterior distribution of $c_i$. The prediction of $X_i$ is $\hat{X}_i = \hat{\mu} + \mathbf{\Psi}\hat{c}_i$. In addition, the asymptotic covariance matrix of $\hat{c}_i$ is

$$\text{cov}(\hat{c}_i) = \hat{\sigma}^2 (\mathbf{\Psi}_p^\top \mathbf{\Psi}_p + \hat{\sigma}^2 \hat{\mathbf{\Gamma}}^{-1})^{-1}, \tag{20}$$

and the asymptotic covariance matrix of $\hat{X}_i$ is given by

$$\text{cov}(\hat{X}_i) = \mathbf{\Psi}_q (N\mathbf{\Psi}_q^\top \hat{\mathbf{\Sigma}}_y^{-1} \mathbf{\Psi}_q + \gamma \mathbf{Q})^{-1} \mathbf{\Psi}_q^\top + \hat{\sigma}^2 \mathbf{\Psi}_p (\mathbf{\Psi}_p^\top \mathbf{\Psi}_p + \hat{\sigma}^2 \hat{\mathbf{\Gamma}}^{-1})^{-1} \mathbf{\Psi}_p^\top, \tag{21}$$

which is used to constructed the confidence band of $\hat{X}_i$.

The novelty of our new estimating procedure is the estimator of $\mathbf{\Gamma}$ and $\sigma^2$. Specifically, our method employs the Laplace Approximation Marginal Likelihood (LAML) [39] to estimate $\mathbf{\Gamma}$ and $\sigma^2$, which has a much lower computational cost than REML. The objective function of $\mathbf{\Gamma}$ and $\sigma^2$ based on LAML is given by

$$\begin{aligned}
\mathcal{L}(\sigma^2, \mathbf{\Gamma}) = \sum_{i=1}^{N} & \left\{ \frac{1}{\hat{\sigma}^2} ||Y_i - \hat{\mu} - \mathbf{\Psi}_p \hat{c}_i||_2^2 + \hat{c}_j^\top \mathbf{\Gamma}^{-1} \hat{c}_j + \log\det\mathbf{\Gamma}^{-1} + \log(\sigma^2 \mathbf{I}_T) \right\} \\
& + \log\det\left( \frac{\mathbf{\Psi}_p^\top \mathbf{\Psi}_p}{\sigma^2} + \mathbf{\Gamma}^{-1} \right).
\end{aligned} \tag{22}$$

In this objective function, the first four components come from the joint log-likelihood function of $\{Y_i\}$ and $\{c_i\}$, while the fifth term is resulted by the Laplace approximation of integral $\int f(Y_i|c_i) f(c_i) dc_i$, where $f(Y_i|c_i)$ is the density function of $Y_i$ conditioned on $c_i$ and $f(c_i)$ is the density function of $c_i$. The derivative of $\mathcal{L}(\sigma^2, \mathbf{\Gamma})$ with respective to $\sigma^{-2}$ is

$$\frac{\partial \mathcal{L}(\sigma^2, \mathbf{\Gamma})}{\partial \sigma^{-2}} = \mathbf{0} = -NT\sigma^2 + \sum_{i=1}^{N} \left( ||Y_i - \hat{\mu} - \mathbf{\Psi}_p \hat{c}_i||_2^2 \right) + \text{trace}(\mathbf{\Delta}(\sigma^2, \mathbf{\Gamma}) \mathbf{\Psi}_p^\top \mathbf{\Psi}_p), \tag{23}$$

where $\mathbf{\Delta}(\sigma^2, \mathbf{\Gamma}) = (\mathbf{\Psi}_p^\top \mathbf{\Psi}_p / \sigma^2 + \mathbf{\Gamma}^{-1})^{-1}$. As a result, the fixed-point iteration for $\sigma^2$ is

$$\sigma^{2[t+1]} = \frac{1}{NT} \left\{ \sum_{i=1}^{N} \left( ||Y_i - \hat{\mu} - \mathbf{\Psi}_p \hat{c}_i||_2^2 \right) + \text{trace}(\mathbf{\Delta}(\sigma^{2[t]}, \mathbf{\Gamma}^{[t]}) \mathbf{\Psi}_p^\top \mathbf{\Psi}_p) \right\}, \tag{24}$$

where $\sigma^{2[t]}$ and $\mathbf{\Gamma}^{[t]}$ are the current estimates. On the other hand, the derivative of $\mathcal{L}(\sigma^2, \mathbf{\Gamma})$ with respective to $\mathbf{\Gamma}^{-1}$ is

$$\frac{\partial \mathcal{L}(\sigma^2, \mathbf{\Gamma})}{\partial \mathbf{\Gamma}^{-1}} = -N\mathbf{\Gamma} + \sum_{i=1}^{N} \left( \hat{c}_i \hat{c}_i^\top \right) + \mathbf{\Delta}(\sigma^2, \mathbf{\Gamma}). \tag{25}$$

Hence, the fixed-point iteration for $\boldsymbol{\Gamma}$ is

$$\boldsymbol{\Gamma}^{[t+1]} = \frac{1}{N}\left\{ \sum_{i=1}^{N}\left(\hat{\boldsymbol{c}}_i\hat{\boldsymbol{c}}_i^{\top}\right) + \boldsymbol{\Delta}(\sigma^{2[t]},\boldsymbol{\Gamma}^{[t]}) \right\}. \tag{26}$$

We implement these two fixed-point iterations iteratively and regard the stable solutions as $\hat{\sigma}^2$ and $\hat{\boldsymbol{\Gamma}}$. Likewise, we also implement the estimator of $\boldsymbol{\beta}$, the predictors of $\{\boldsymbol{c}_i\}$, and the estimator of $\boldsymbol{\Gamma}$ and $\sigma^2$ iteratively, and regard the stable solutions as the outputting estimates of $\boldsymbol{\beta}$, $\{\boldsymbol{c}_i\}$, $\boldsymbol{\Gamma}$, and $\sigma^2$.

The reason why LAML is more efficient than REML is illustrated as follows. The objective function of REML with respect to balanced longitudinal data is

$$\mathcal{R}(\boldsymbol{\vartheta}) = \sum_{i=1}^{N}\left( (\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\hat{\boldsymbol{\beta}})^{\top}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\hat{\boldsymbol{\beta}}) + \log\det(\boldsymbol{\Sigma}_y) \right) + \log\det(\boldsymbol{\Psi}_q^{\top}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Psi}_q), \tag{27}$$

where $\boldsymbol{\vartheta} = (\sigma^2, \text{vec}(\boldsymbol{\Gamma}))$ and $\boldsymbol{\Sigma}_y \approx \sigma^2\boldsymbol{I}_T + \boldsymbol{\Psi}_p\boldsymbol{\Gamma}\boldsymbol{\Psi}_p^{\top}$. Because $\sigma^2$ and $\boldsymbol{\Gamma}$ are involved in $\boldsymbol{\Sigma}_y$, it is impossible to obtain fixed-point iterations for $\sigma^2$ and $\boldsymbol{\Gamma}$ separately as LAML. In other words, it seems that the only way to jointly estimate $\sigma^2$ and $\text{vec}(\boldsymbol{\Gamma})$ is by using the Fisher-scoring algorithm. The first-order derivative of $\mathcal{R}(\boldsymbol{\vartheta})$ with respect to $\vartheta_k$ is

$$\frac{\partial\mathcal{R}(\boldsymbol{\vartheta})}{\partial\vartheta_k} = 0 = -\sum_{i=1}^{N}(\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\hat{\boldsymbol{\beta}})^{\top}\boldsymbol{\Sigma}_y^{-1}\frac{\partial\boldsymbol{\Sigma}_y}{\partial\vartheta_k}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\hat{\boldsymbol{\beta}}) + \text{trace}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Sigma}_y}{\partial\vartheta_k}\right), \tag{28}$$

while the expectation of second-order derivative of $\mathcal{R}(\boldsymbol{\vartheta})$ with respect to $\vartheta_k$ and $\vartheta_h$, i.e., the $(k,h)$th element of the Hessian matrix of $\mathcal{R}(\boldsymbol{\vartheta})$, is given by

$$\mathrm{E}\left(\frac{\partial^2\mathcal{R}(\boldsymbol{\vartheta})}{\partial\vartheta_k\partial\vartheta_h}\right) = \text{trace}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\Sigma}_y}{\partial\vartheta_k}\boldsymbol{P}\frac{\partial\boldsymbol{\Sigma}_y}{\partial\vartheta_h}\right), \tag{29}$$

where

$$\boldsymbol{P} = \boldsymbol{\Sigma}_y^{-1} - \boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Psi}_q(\boldsymbol{\Psi}_q^{\top}\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Psi}_q)^{-1}\boldsymbol{\Psi}_q^{\top}\boldsymbol{\Sigma}_y^{-1}. \tag{30}$$

Thus, $\boldsymbol{\vartheta}$ is updated by

$$\boldsymbol{\vartheta}^{[t+1]} = \boldsymbol{\vartheta}^{[t]} - \boldsymbol{H}(\boldsymbol{\vartheta}^{[t]})^{-1}\boldsymbol{g}(\boldsymbol{\vartheta}^{[t]}), \tag{31}$$

where the $k$th element of $\boldsymbol{g}(\boldsymbol{\vartheta}^{[t]})$ is calculated according to (28) and the $(k,h)$th element of $\boldsymbol{H}(\boldsymbol{\vartheta}^{[t]})$ is yielded by (29).

It is worth pointing out that there is no matrix formula to directly yield $\boldsymbol{g}(\boldsymbol{\vartheta}^{[t]})$ and $\boldsymbol{H}(\boldsymbol{\vartheta}^{[t]})$, i.e., it is the sole way to calculate their elements one-by-one according to (28) and (29). This drawback makes the minimization of REML extremely costly because the computation complexity of matrix inverse $\boldsymbol{\Sigma}_y^{-1}$ and matrix product $\boldsymbol{P}\partial\boldsymbol{\Sigma}_y/\partial\vartheta_k$ are of $O(T^3)$. In particular, as the number of employed basis function $p$ increases, the number of elements in $\boldsymbol{H}(\boldsymbol{\vartheta}^{[t]})$ is $(p^2+1)^2$, which will be more than $10,000$ if $p$ is just 10. As a result, the total computation complexity of the Hessian matrix calculation is $O(T^3p^4)$, which is unacceptable for large balanced longitudinal data. In contrast, our separate fixed-point iterations (24) and (26) have only $O(Tp^2)$ computation complexity, and are able to guarantee $\hat{\sigma}^2$ positive and $\hat{\boldsymbol{\Gamma}}$ positive definite as long as the initial estimates $\sigma^{2[0]}$ is positive and $\boldsymbol{\Gamma}^{[0]}$ is positive definite. Therefore, our novel estimator is more attractive than the REML-based one for FLMM in high dimensionality. Note that what we contribute is the fixed-point iterations based on LAML. In the paper of LAML [39], the Newton–Raphson algorithm was employed to minimize the corresponding LAML objective function, in which will still face the same numerical problems as REML if $T$ or $p$ is large.

### *3.2. Regularity Assumptions*

Now we turn to investigate the large sample properties of FLMM using our novel estimator. We first give three essential concepts that play central roles in our asymptotic investigation.

**Definition 1** (Sub-Gaussianity). *A random error $\varepsilon_i$ with mean $\mathrm{E}(\varepsilon_i) = 0$ and variance $var(\varepsilon_i) = \sigma^2$ is termed a sub-Gaussian variable if for all $r \in \mathrm{R}$,*

$$\mathrm{E}(\exp(r\varepsilon_i)) \leq \exp\left(\frac{1}{2}\tau^2 r^2\right),$$

*where $\tau > 0$ is called the sub-Gaussian parameter.*

**Definition 2** (Well-conditioned covariance matrix). *An $(p \times p)$ symmetric matrix $\mathbf{\Sigma}$ is termed the well-conditioned covariance matrix if there is a constant $c_0$ independent of $p$ such that*

$$0 < c_0^{-1} \leq \lambda_{\min}(\mathbf{\Sigma}) \leq \lambda_{\max}(\mathbf{\Sigma}) \leq c_0 < \infty.$$

**Definition 3** (Sobolev–Hilbert space). *A functional space $\mathcal{F}_s$ is called s-order Sobolev–Hilbert space if there is a constant $c_0$ such that $\forall f \in \mathcal{F}_s$,*

$$E\left(\int_{t \in \mathcal{T}} [f^s(t)]^2 dt\right) \leq c_0 < \infty.$$

Sub-Gaussianity is an important concept in high-dimensional statistical analysis, which generalizes the Gaussian distribution to include common continuous variables and all bounded discrete variables [40]. Besides, the concept of well-conditioned covariance is proposed by Bickel and Levina [41], which guarantees that all of its eigenvalues are bounded away from zero and infinity whatever its dimension is. In addition, Sobolev–Hilbert space is the most commonly-used functional space in nonparametric statistical analysis [42]. The minimax lower bound of function estimate in this space is

$$\inf_{\hat{f}} \sup_{f \in \mathcal{F}_s} \mathrm{E}(||\hat{f} - f||_2^2) \geq d_0 N^{-\frac{2s}{2s+1}}, \tag{32}$$

where $d_0 > 0$ is a certain constant and $N$ describes the sample size.

Now we give the following assumptions that facilitate the proofs of theorems.

**Assumption 1** (Sub-Gaussian random error). *For all $(i, j) \in \{1, \ldots, N\} \times \{1, \ldots, T\}$, the random error $\varepsilon_{ij}$ is independently and identically distributively generated from a sub-Gaussian distribution with mean 0, variance $\sigma^2$, and sub-Gaussian parameter $\tau_\epsilon$.*

**Assumption 2** (Quasi uniform locations). *The sampling locations $t_1, \ldots, t_T$ form a quasi uniform sequence defined in $\mathcal{T}$.*

**Assumption 3** (Basis functions). *For a series of basis functions $\{\Psi_k(t)\}$,*

$$\int_{t \in \mathcal{T}} \Psi_k(t) d_t = 0, \quad \int_{t \in \mathcal{T}} [\Psi_k(t)]^2 dt = 1, \quad \int_{t \in \mathcal{T}} [\Psi_k(t)]^4 dt \leq c_0 < \infty,$$

*for all $k \in \{1, \ldots, p\}$, where $c_0 > 0$ is a constant. Besides, let $\mathbf{\Psi}_p(t) = (\Psi_1(t), \ldots, \Psi_p(t))^\top$ and*

$$\mathbf{\Sigma}_{\Psi_p} = \int_{t \in \mathcal{T}} \mathbf{\Psi}_p(t) \mathbf{\Psi}_p(t)^\top dt.$$

*Then $\mathbf{\Sigma}_{\Psi_p}$ is a well-conditioned covariance matrix.*

**Assumption 4** (Approximation error of mean function). *For a mean function $\mu(t) \in \mathcal{F}_s$ and a series of basis functions $\{\Psi_k(t)\}$ satisfying Assumption 3, there exists a unique series of coefficients $\{\beta_k\}$ satisfying $\lim_{q \to \infty} \sum_{k=1}^{q} \beta_k^2 \leq c_0 < \infty$ such that*

$$\int_{t \in \mathcal{T}} \left( \mu(t) - \sum_{k=1}^{q} \Psi_k(t)\beta_k \right)^2 dt \leq c_0 q^{-2s},$$

*where $c_0 > 0$ is certain constant.*

**Assumption 5** (Approximation error of individual function). *For any individual trajectory $X_i(t) \in \mathcal{F}_s$ and a series of basis functions $\{\Psi_k(t)\}$ satisfying Assumption 3, there exists a unique series of random coefficients $\{c_{ik}\}$ satisfying $\lim_{p \to \infty} \sum_{k=1}^{p} E(c_{ik}^2) \leq c_0 < \infty$ such that*

$$E \left( \int_{t \in \mathcal{T}} \left[ X_i(t) - \mu(t) - \sum_{k=1}^{p} \Psi_k(t)c_{ik} \right]^2 dt \right) \leq c_0 p^{-2s}, \quad i \in \{1, \ldots, N\},$$

*where $c_0 > 0$ is certain constant.*

**Assumption 6** (Distribution of random coefficients). *Random coefficient $c_{ik}$ is sub-Gaussian with $E(c_{ik}) = 0$, $var(c_{ik}) = \Gamma_{kk}$, sub-Gaussian parameter $\tau_k$, and $c_{ik}$ is independent of $c_{jk}$ for all $i \neq j$. Furthermore, $T\Gamma = (T\Gamma_{jk})_{p \times p}$ is a well-conditioned covariance matrix.*

Assumption 1 considers the random error to be independent and identically distributed (IID) sub-Gaussian variable for all $i$ and $j$. Assumption 2 sets the sampling locations and $t_1, \ldots, t_T$ can be regarded as a uniform sequence. As a result, the related integral of a certain function of $t$ can be approximated by averaging the values of this function at these discrete locations. Assumption 3 summarizes some basic properties of the basis functions $\{\Psi_k(t)\}$. Assumptions 4 and 5 describe how the accuracy of a function can be approximated by using $p$ bases. Tsybakov [43] showed that, if the bases are the Tikhonov bases, Assumptions 4 and 5 hold as long as $\mu(t)$ and $X_i(t)$ belong to the Sobolev–Hilbert space $\mathcal{F}_s$. However, for a series of general basis functions, it remains unclear if Assumptions 4 and 5 still hold. In other words, we assume that we have picked the basis functions that have similar properties as the Tikhonov bases such that these two hold. Assumption 6 lists some conditions of the random coefficients. In particular, since

$$\lambda_{\max}(\Psi \Gamma \Psi^\top) = O_P(T) \times \lambda_{\max}(\Sigma_\Psi) \times \lambda_{\max}(\Gamma), \tag{33}$$

$\lambda_{\max}(\Gamma)$ must be $O(T^{-1})$ otherwise the covariance matrix of $\mathbf{Y}_i$ becomes divergent and violates some basic assumption of FDA; see, e.g., assumption (A4) in Yao et al. [11]. Hence, we assume that the covariance matrix of $\sqrt{T}c_i$, i.e., $T\Gamma$, is well-conditioned.

### 3.3. Large Sample Property

We provide four theorems describing the asymptotic properties of estimators obtained by our new FLMM.

**Theorem 1.** *Suppose Assumptions 1–6 hold. If $\hat{\beta}$ is yielded by (14), $\lambda_{\max}(\gamma \mathbf{Q}) = O(1)$, and $\hat{\Sigma}_y$ is chosen such that $\lambda_{\max}(\hat{\Sigma}_y^{-1} \Sigma_y) \leq c_0 < \infty$ for certain constant $c_0 > 0$, then*

$$||\hat{\beta} - \beta||_2^2 = O_P \left( \frac{q}{NT} + \frac{1}{q^{2s}} \right).$$

*If $q = O(T)$, then the optimal convergence rate of $\hat{\beta}$ is*

$$||\hat{\beta} - \beta||_2^2 = O_P \left( \frac{1}{N} + \frac{1}{T^{2s}} \right).$$

Theorem 1 indicates the convergence rate of the mean parameter estimate $\hat{\boldsymbol{\beta}}$, which is a trade-off between variance term $O_P(q/NT)$ and bias term $O(p^{-2s})$. As a corollary,

$$
\int_{t \in \mathcal{T}} (\mu(t) - \hat{\mu}(t))^2 dt \leq \int_{t \in \mathcal{T}} (\mu(t) - \boldsymbol{\Psi}_q(t)^\top \boldsymbol{\beta})^2 dt + \frac{1}{T} \sum_{j=1}^{T} (\boldsymbol{\Psi}_q(t_j)^\top \{\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\})^2
$$
$$
= O_P\left( \frac{q}{NT} + \frac{1}{q^{2s}} \right).
$$
(34)

which means that the optimal convergence rate $\hat{\mu}(t)$ is $O_P(N^{-1} + T^{-2s})$ by letting $q = O(T)$. In particular, this convergence rate reaches the minimax lower bound of all nonparametric estimates of $\mu(t)$ [29,43], indicating that our estimate is optimal in terms of minimax efficiency. Besides, our theory does not require $\lambda$ to vanish asymptotically. Indeed, any $\gamma$ such that $\lambda_{\max}(\gamma \mathbf{Q}) = O(1)$ is able to guarantee $\hat{\boldsymbol{\beta}}$ to reach the minimax lower bound.

**Theorem 2.** *Suppose Assumptions 1–6 hold. If $\hat{\boldsymbol{c}}_i$ is yielded by (19), then for $i = 1, \ldots, N$,*

$$
||\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i||_2^2 = O_P\left( \frac{p}{T} + \frac{1}{p^{2s}} \right).
$$

*If $p = O(T^{\frac{1}{2s+1}})$, then the optimal convergence rate of $\hat{\boldsymbol{c}}_i$ is*

$$
||\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i||_2^2 = O_P\left( T^{-\frac{2s}{2s+1}} \right).
$$

Theorem 2 shows the convergence rate of the parameter estimate in individual trajectory $\hat{\boldsymbol{c}}_i$, which is a trade-off between variance term $O_P(p/T)$ and bias term $O(p^{-2s})$. Similarly, the optimal convergence rate of $\hat{X}_i(t)$ is

$$
\int_{t \in \mathcal{T}} (X_i(t) - \hat{X}_i(t))^2 dt \leq \int_{t \in \mathcal{T}} (X_i(t) - \mu(t) - \boldsymbol{\Psi}_p(t)^\top \boldsymbol{c}_i)^2 dt + \int_{t \in \mathcal{T}} (\mu(t) - \hat{\mu}(t))^2 dt
$$
$$
+ \frac{1}{T} \sum_{j=1}^{T} (\boldsymbol{\Psi}_p(t_j)^\top \{\boldsymbol{c}_i - \hat{\boldsymbol{c}}_i\})^2 = O_P\left( \frac{p}{T} + \frac{1}{p^{2s}} \right),
$$
(35)

which reaches the minimax lower bound of all nonparametric estimates of $X_i(t)$ [29,43] by letting $p = O(T^{1/(2s+1)})$. In addition, it is easy to see that the optimal $q$ and the optimal $p$ have obviously different orders of magnitude: the former should diverge at the same rate as $T$ while the latter is much slower than $T$. Hence, to reach the optimal convergence rates, we should use an adaptive number of bases when estimating $\mu(t)$ and predicting $X_i(t)$. As far as we are concerned, we are the first ones to indicate this principle. Traditional FLMM methods, such as James et al. [17] and Shi et al. [20], did not require different numbers of bases when estimating the mean function and individual function, which may result in less efficient estimates of mean function and individual function in practice.

**Theorem 3.** *Suppose Assumptions 1–6 hold. If $\hat{\sigma}^2$ is yielded by (24),*

$$
(\hat{\sigma}^2 - \sigma^2)^2 = O_P\left( \frac{p}{T} + \frac{1}{p^{2s}} \right).
$$

*If $p = O(T^{\frac{1}{2s+1}})$, the optimal convergence rate is*

$$
(\hat{\sigma}^2 - \sigma^2)^2 = O_P\left( T^{-\frac{2s}{2s+1}} \right).
$$

**Theorem 4.** *Suppose Assumptions 1–6 hold. If $\hat{\boldsymbol{\Gamma}}$ is yielded by (26),*

$$||\boldsymbol{\Gamma}^{-\frac{1}{2}}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^{-\frac{1}{2}} - \mathbf{I}_p||_2^2 = O_P\left(\frac{p}{N} + \frac{p}{T} + \frac{1}{p^{2s}}\right).$$

*If $p = O(T^{\frac{1}{2s+1}})$, the optimal convergence rate is*

$$||\boldsymbol{\Gamma}^{-\frac{1}{2}}\hat{\boldsymbol{\Gamma}}\boldsymbol{\Gamma}^{-\frac{1}{2}} - \mathbf{I}_p||_2^2 = O_P\left\{ \max\left( N^{-1}T^{\frac{1}{2s+1}}, T^{-\frac{2s}{2s+1}} \right) \right\}.$$

Theorems 3 and 4 guarantee that the variance components $\sigma^2$ and $\boldsymbol{\Gamma}$ can be estimated consistently by using the LAML criterion. The convergence rates of $\sigma^2$ and $\boldsymbol{\Gamma}$ will be influenced by the estimation errors of $\hat{\boldsymbol{\beta}}$ and $\hat{c}_i$. As a corollary,

$$||\boldsymbol{\Sigma}_y^* - \hat{\boldsymbol{\Sigma}}_y||_2^2 = O_P\left( N^{-1}T^{\frac{1}{2s+1}} + T^{-\frac{2s}{2s+1}} \right), \tag{36}$$

where $\boldsymbol{\Sigma}_y^* = \sigma^2\mathbf{I}_T + \boldsymbol{\Psi}_p\boldsymbol{\Gamma}\boldsymbol{\Psi}_p^\top$ and $\hat{\boldsymbol{\Sigma}}_y = \hat{\sigma}^2\mathbf{I}_T + \boldsymbol{\Psi}_p\hat{\boldsymbol{\Gamma}}\boldsymbol{\Psi}_p^\top$. However, the true covariance matrix is $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{I}_T + \text{cov}(\boldsymbol{X}_i)$. The difference between $\boldsymbol{\Sigma}_y$ and $\boldsymbol{\Sigma}_y^*$ is unknown without additional assumption. Hence, in Theorem 1, we give the condition $\lambda_{\max}(\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Sigma}_y) < \infty$ to ensure the correctness of the corresponding proof.

*3.4. Tuning Parameter Selection*

In the new estimator of FLMM, three tuning parameters including $p$, $q$, and $\gamma$ need to be determined artificially. Here, we should emphasize that our FLMM estimator is insensitive to the choices of $p$, $q$ as long as they are moderately large (e.g., $p = 10$ and $q = 20$, or any other reasonable numbers), though the theoretical investigation requires them to be bounded by $O(T^{1/(2s+1)})$ and $O(T)$. In terms of regulating the wigglinesses of the mean function estimate and individual trajectory prediction, the smoothing parameter $\gamma$ and covariance matrix $\boldsymbol{\Gamma}$ are more essential than the numbers of basis functions $p$ and $q$. A large number of empirical analyses have confirmed this principle. For example, Wood [44] showed it can accurately fit a common function by setting $p = 10$ and set this number as the default one in R package mgcv. We follow Wood's measure and recommend choosing $p$ and $q$ as any reasonable numbers in this paper.

As for $\gamma$, we propose to automatically estimate the optimal one by using the LAML:

$$\hat{\gamma} = \arg\min_\gamma \left\{ \gamma\hat{\boldsymbol{\beta}}^\top\mathbf{Q}\hat{\boldsymbol{\beta}} - \log\det(\gamma\mathbf{I}_{\text{rank}(\mathbf{Q})}) + \log\det\left( n\boldsymbol{\Psi}_q^\top\boldsymbol{\Psi}_q + \gamma\mathbf{Q} \right) \right\}. \tag{37}$$

Similar to the estimation of $\boldsymbol{\Gamma}$ and $\sigma^2$, the score function of $\gamma$ is

$$0 = \hat{\boldsymbol{\beta}}^\top\mathbf{Q}\hat{\boldsymbol{\beta}} - \text{rank}(\mathbf{Q})\gamma^{-1} + \text{trace}((n\boldsymbol{\Psi}_q^\top\boldsymbol{\Psi}_q + \gamma\mathbf{Q})^{-1}\mathbf{Q}). \tag{38}$$

As a result, the fixed-point iteration of $\gamma$ is

$$\gamma^{[t+1]} = \frac{\text{rank}(\mathbf{Q})}{\hat{\boldsymbol{\beta}}^\top\mathbf{Q}\hat{\boldsymbol{\beta}} + \text{trace}((n\boldsymbol{\Psi}_q^\top\boldsymbol{\Psi}_q + \gamma^{[t]}\mathbf{Q})^{-1}\mathbf{Q})}. \tag{39}$$

We update $\gamma^{[t]}$ along with the estimation of $\boldsymbol{\beta}$. The stable estimate is regarded as $\hat{\gamma}$.

Since $\hat{\gamma}$ can be estimated automatically, there is indeed no tuning parameter to be determined. In contrast, the estimating procedure of FPCA needs to select three bandwidths

and the number of employed PCs, which is supposedly very time-consuming. In particular, FPCA is sensitive to the number of employed PCS, while FLMM is insensitive to numbers of the employed bases $p$ and $q$. As a result, our new estimator of FLMM outperforms FPCA absolutely in terms of tuning-parameter robustness.

## 4. Simulation Study

In this section, we conduct a simulation study to assess the performance of our new FLMM estimator in comparison with FPCA. Here, we choose the well-known Principal Analysis by Conditional expectation (PACE) estimator (Matlab codes of PACE: https://www.stat.ucdavis.edu/PACE/, from 16 March 2020 to 14 August 2020) to estimate the FPCA.

### 4.1. Simulation Settings

The settings of simulations are as follows. Throughout the study, we set the $t = (1/T, 2/T, \cdots, 1)^\top$ to keep the $t \in (0, 1)$. The mean function is

$$\mu(t) = t^{11}(10(1-t))^6 + 10(10t)^3(1 - t^{11}) - 1.396, \tag{40}$$

which is constructed by Wood [44]. Next, we generate the covariance function $C(s, t)$ by

$$C(s, t) = \sum_{k=1}^{5} \lambda_k \phi_k(s) \phi_k(t). \tag{41}$$

Here, the eigenvalues $(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = (5, 2.5, 30, 0.5, 0.25)/4$, and the eigenfunction vector $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_5(t))^\top$ is constructed by rotating the Tikhonov bases through a $(5 \times 5)$ matrix $\mathbf{G}$:

$$\boldsymbol{\phi}(t) = \boldsymbol{T}(t)\mathbf{G}, \tag{42}$$

where $\boldsymbol{T}(t) = (1, \sqrt{2}\cos(\pi t), \ldots, \sqrt{2}\cos(4\pi t))^\top$, $\mathbf{A}_5(0.5)$ is a $(5 \times 5)$ 1-order autoregressive (AR(1)) structure correlation matrix with correlation coefficient $\rho = 0.5$, and $\mathbf{G}$ is the Cholesky factor of $\mathbf{A}_5(0.5)^{-1}$, i.e., $\mathbf{G}^\top\mathbf{G} = \mathbf{A}_5(0.5)^{-1}$. Furthermore, the individual trajectory is generated by

$$X_i(t_j) = \mu(t_j) + \sum_{k=1}^{5} \phi_k(t_j)\xi_{ik}, \quad (i, j) \in \{1, \ldots, N\} \times \{1, \ldots, T\}, \tag{43}$$

where $\xi_{ik} \sim \mathcal{N}(0, \lambda_k)$. The observation

$$y_i(t_j) = X_i(t_j) + \varepsilon_{ij}, \tag{44}$$

where $\varepsilon_{ij} \sim N(0, 2)$. Figure 2 visualizes the aforementioned settings. Specifically, the left-top panel shows the first three Tikhonov bases and the left-bottom panel demonstrates the first three eigenfunctions rotated by the Tikhonov bases. The middle-top and middle-bottom panels show the first and second individual trajectories and the related observations with random errors. It is observed that the individual trajectory is similar to the mean function but has its adaptive trend. The right-top and right-bottom panels visualize the covariance surfaces.
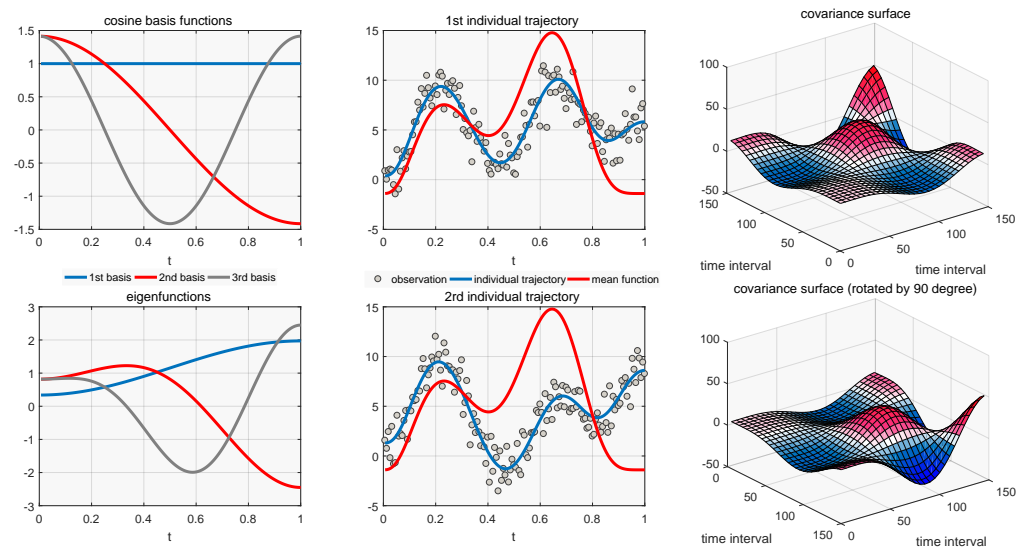
**Figure 2.** The eigenfunctions, mean function, individual trajectory, and covariance function used in the simulation.

### 4.2. Simulation Results

For better comparison, we notated FLMM with Bsplines and Tikhonov bases as FLMM-Bsplines and FLMM-Tikhonov, respectively. The numbers of employed bases for Bsplines and Tikhonov bases are 30 and 15 when estimating the mean function trajectory and individual trajectory, respectively. Meanwhile, the FPCA implemented by PACE is noted as FPCA-PACE, where all settings follow the default ones in the package. We compare the proposed FLMM with FPCA and generate the data using the above settings. In order the get the whole evaluation of the three approaches, we take two cases in the simulations: (1) fix the $N$ as 150, and set $T = 50, 100, 150, 200, 250, 300$; (2) fix $T = 150$, set $N = 50, 100, 150, 200, 250, 300$. On the other hand, the criteria to evaluate the results are as follows. For mean function estimation, we adopt the root mean squared error (RMSE) as the criterion:

$$\sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(\hat{\mu}(t_i) - \mu(t_i)\right)^2} \tag{45}$$

for individual trajectory prediction, we consider the maximum RMSE as the criterion:

$$\max_{i\in\{1,\dots,N\}}\sqrt{\frac{1}{T}\sum_{i=1}^{T}\left(\hat{X}_i(t_i) - X_i(t_i)\right)^2} \tag{46}$$

the absolute value $|\hat{\sigma}^2 - \sigma|$ is used to evaluate the variance estimation; the Frobenius-norm of the matrix is adopted as a measure to compare the covariance surface estimation:

$$\sqrt{\frac{1}{T}\sum_{i=1}^{T}\sum_{j=1}^{T}\left(\hat{C}(t_i, t_j) - C(t_i, t_j)\right)^2} \tag{47}$$

and the computing time in seconds is recorded to compare the computational efficiency. The system and software to implement the simulations are Linux 4.18 (standard High Performance Computing (HPC) machine, icosa192gb feature nodes, memory 50 GB and 12 cpu-cores) and MATLAB (R2021). The replication runs are 500.

Figure 3 shows the results of the simulation. Specifically, the five subplots on the left side of Figure 3 show the results of FLMM-Bsplines, FLMM-Tikhonov, and FPCA-PACE based on mean function estimation, individual estimation prediction, variance

estimation, covariance surface estimation, and computation time as the sample size *N* increases. Likewise, the five subplots on the right side of Figure 3 show the counterparts of the left subplots as the number of observations *T* increases. Based on these results, we make the following comments.
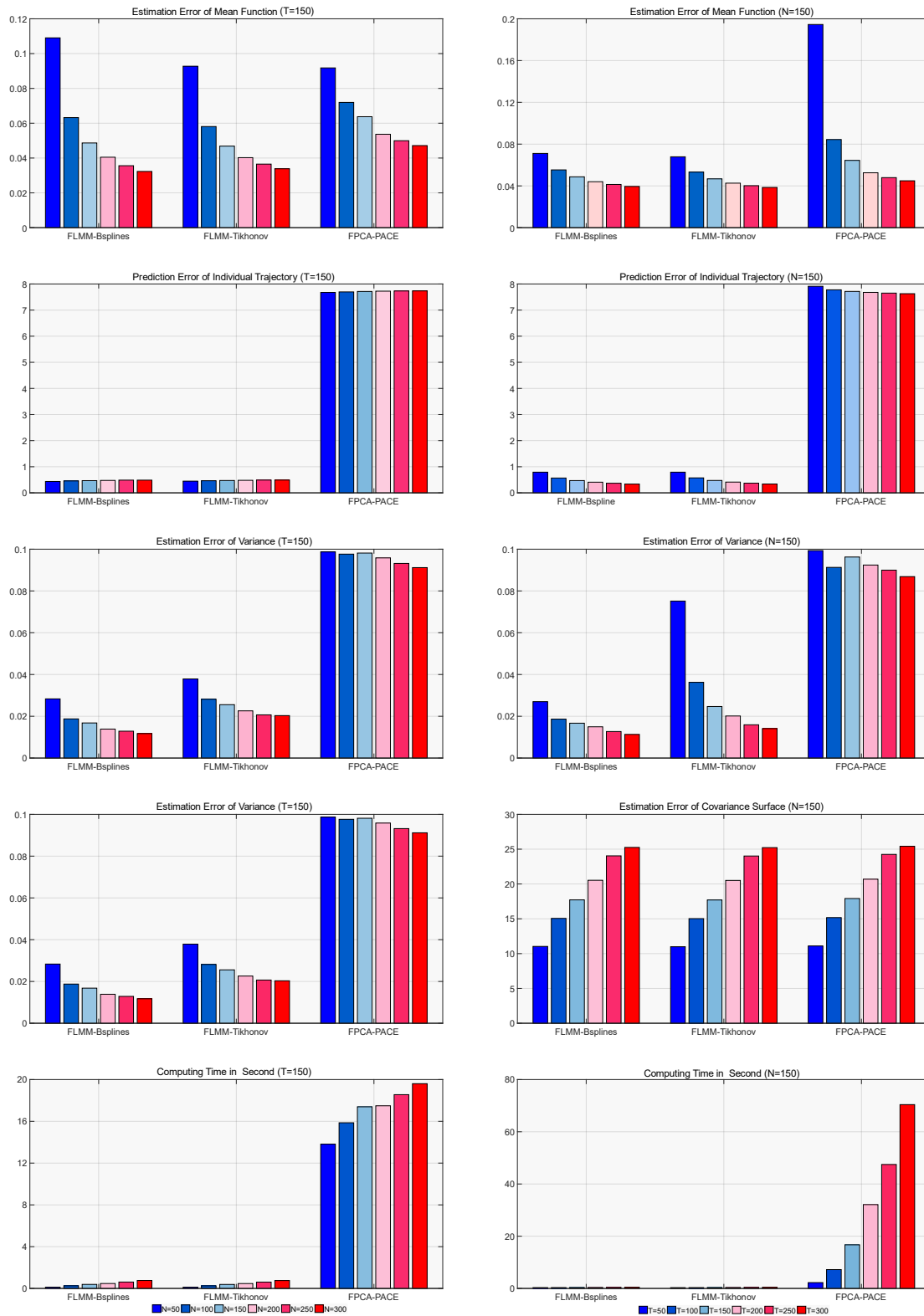


**Figure 3.** Estimation errors of mean function, individual trajectory, variance, covariance surface, and running time in the three methods.

*Mean function estimation.* When sample size $N$ is very small, FLMM-Bsplines is worse than FLMM-Tikhonov and FPCA-PACE. However, as $N$ increases, the performances of FLMM-Bsplines and FLMM-Tikhonov boost significantly faster than FPCA-PACE. On the other hand, if $N$ is fixed to 150, FPCA-PACE is extremely unreliable when $T$ is very small and is always less accurate than them no matter whether $T$ is large or small. These results show that our novel FLMM estimator is better than FPCA in terms of mean function estimation. In addition, no matter whether $N$ or $T$ rises, the estimation error of the mean function will decrease, which is consistent with Theorem 1.

*Individual trajectory prediction.* In terms of individual trajectory prediction, FPCA-PACE is much worse than FLMM-Bsplines and FLMM-Tikhonov. We checked the codes of the PACE package and found that it estimates the eigenvalues eigenfunctions by applying PCA on a $(51 \times 51)$ covariance surface predictor, which may result in less accurate eigenvalues eigenfunctions estimates. (Nevertheless, we did not change the default settings of the PACE package in the simulations.) On the other hand, we find that our estimates become more accurate as $T$ increases, while slowly becoming worse when $T$ remains and $N$ rises (the maximum RMSE (46) is affected by $N$). This is consistent with Theorem 2: the prediction of individual trajectory is only related to the number of observations $T$.

*Variance estimation.* Regarding variance estimation, FLMM-Bsplines and FLMM-Tikhonov enjoy the same accuracy in all cases. Since the variance estimate yielded by FLMM becomes more accurate no matter whether $T$ or $N$ increases, Theorem 3 is supposedly established correctly. However, it seems that FPCA-PACE cannot benefit from either the increase of $T$ or the rise of $N$.

*Covariance surface estimation.* Regarding covariance surface estimation, all three estimators have the same degree of accuracy. In particular, as $N$ is fixed and $T$ increases, all the covariance surface estimators become worse in terms of the Frobenius-norm. This phenomenon indicates that the averaged Frobenius-norm of $||\text{cov}(\boldsymbol{X}_i) - \boldsymbol{\Psi}_p \boldsymbol{\Gamma} \boldsymbol{\Psi}_p^\top||_\text{F}/\sqrt{T}$ should diverge to infinity as $T$ increases.

*Covariance surface approximation.* To verify this hypothesis, we conduct an additional simulation. Specifically, because $\boldsymbol{c}_i$ is unknown in simulation (we generate the random coordinates of eigenfunctions $\boldsymbol{\xi}_i$ directly and there should exist a correspondence between $\boldsymbol{c}_i$ and $\boldsymbol{\xi}_i$), we approximate $\boldsymbol{b}_i$ by

$$\tilde{\boldsymbol{c}}_i = (\boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p)^{-1} \boldsymbol{\Psi}_p^\top (\boldsymbol{X}_i - \boldsymbol{\mu}), \tag{48}$$

and the covariance matrix $\boldsymbol{\Gamma}$ is approximated by

$$\tilde{\boldsymbol{\Gamma}} = \frac{1}{N-1} (\tilde{\boldsymbol{c}}_i - \bar{\boldsymbol{c}})(\tilde{\boldsymbol{c}}_i - \bar{\boldsymbol{c}})^\top, \tag{49}$$

where $\bar{\boldsymbol{c}}$ is the sample mean. We implement the above simulations 300 times, record the empirical covariance matrix $\tilde{\boldsymbol{\Gamma}}$ and the approximation error $||\text{cov}(\boldsymbol{X}_i) - \boldsymbol{\Psi}_p \tilde{\boldsymbol{\Gamma}} \boldsymbol{\Psi}_p^\top||_\text{F}/\sqrt{T}$. The left and middle panels in Figure 4 show the simulated covariance matrices of $\boldsymbol{c}_i$ with respect to Bsplines and Tikhonov bases. Because the basis functions of Bsplines have no natural order while the Tikhonov basis functions have one, the covariance matrix of the Tikhonov basis functions has sharply decreasing diagonal elements with an increasing order. The approximation error does increase with $T$ and so our hypothesis is confirmed. In addition, there is no difference between Bsplines and Tikhonov bases in approximating the covariance surface although the latter have a natural order. In summary, the decreasing performance of covariance surface estimation is due to the approximation error.

*Computational efficiency.* In terms of computational efficiency, the computing time of our two methods is almost independent of $T$ and $N$ and is much less than that of FPCA-PACE. In particular, the computing time of FPCA-PACE increases sharply as $T$ increases, confirming our comments on the drawbacks of FPCA given in Section 2. Thus, for the estimation of large balanced longitudinal data, our proposed method has a very clear advantage.
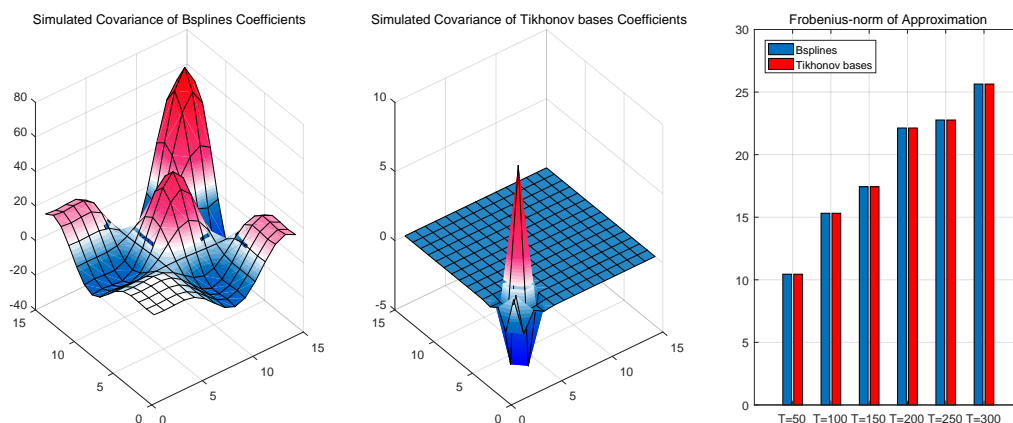
**Figure 4.** Investigation of the covariance surface approximation.

## 5. Real Data Analysis

We demonstrate the effectiveness of the proposed method via the analysis of the COVID-19 data in this section. This data is updated daily on the number of COVID-19 infectious cases and deaths in the 50 states and DC in the US by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. It contains $N = 51$ subjects and $T = 152$ observations. The realization is the number of COVID-19 cases after the logarithm, which was recorded every day from 16 March 2020 to 14 August 2020. Partial observations are negative because of certain unknown reasons. For such an observation, we replace it with the average of the three observations around it. We employ 15 and 10 numbers of Bsplines bases to approximate the mean function and individual trajectories, respectively. It should be pointed out that we have not changed the default settings of the PACE package, and in our FLMM method, the number of employed bases $p$ is the only tuning parameter to select. Therefore, the comparison between FPCA and FLMM is fair.

The results of this data analysis are shown in Figures 5–7. Figure 5 displays the two estimated mean functions obtained by FPCA and our novel FLMM, respectively. The smaller panel within tries to draw attention to the stability of the two mean curves. The results further reveal the reason as to why FPCA performs worse than FLMM in terms of mean function estimation: it may overfit the mean function. Besides, due to the $N = 51$ sample size, we chose six states with the top populations among all states to represent individual trajectories, and the consequence is shown in Figure 6. Likewise, FPCA tends to overfit the individual trajectories while FLMM with our new estimator can fit them accurately. In addition, Figure 7 demonstrated the estimated covariance surface yielded by the two approaches. Generally, the two covariance surfaces are similar; but the one generated by our FLMM is smoother than FPCA. In addition, in terms of computing time, our FLMM estimator only takes 0.088 seconds, while PACE takes 44.08 seconds. Such a substantial gap confirms that our FLMM estimator is much more efficient than FPCA in terms of computation. In conclusion, FPCA implemented by PACE is very likely to overfit the mean function estimation, individual trajectory prediction, and covariance function estimation.
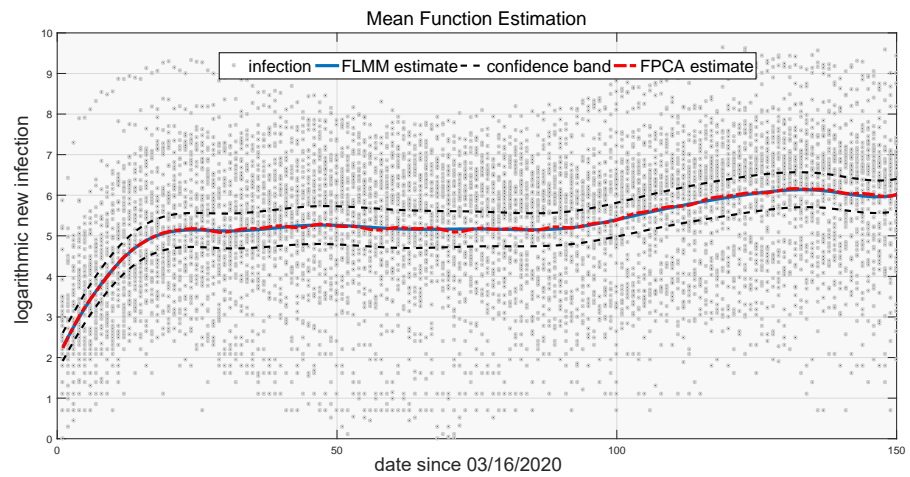
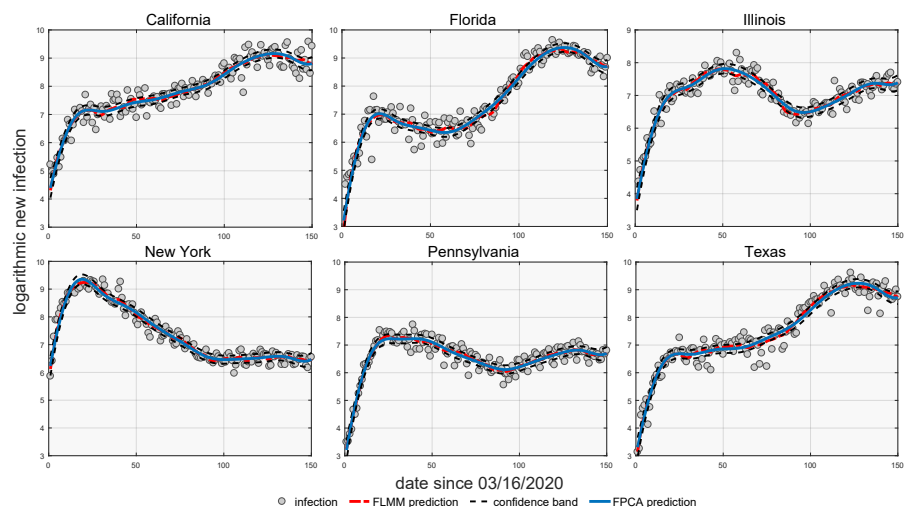**Figure 5.** Fitted mean functions of daily infection via FPCA and proposed FLMM.



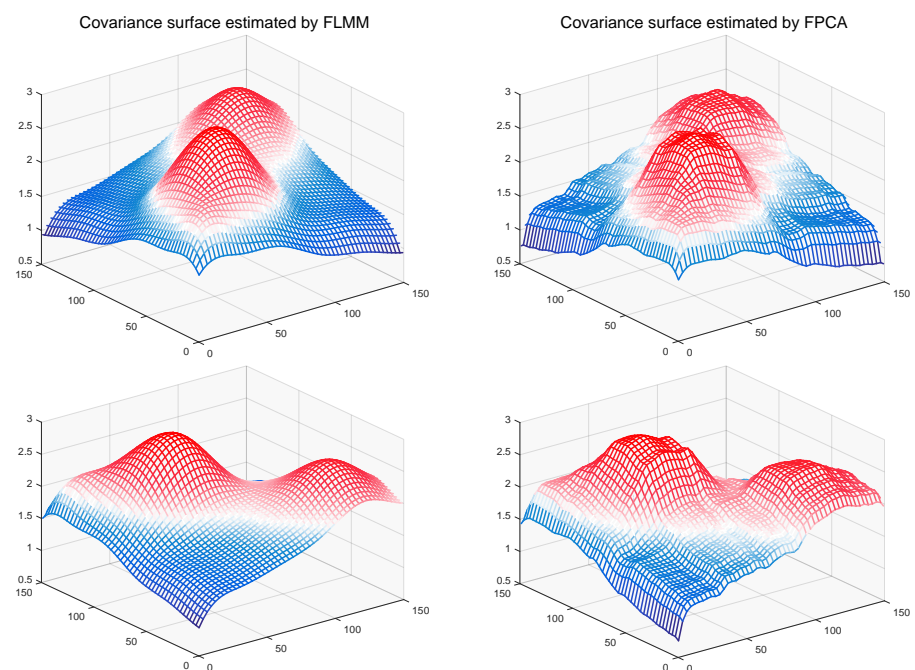**Figure 6.** Predictions of the trajectories of six states via FPCA and proposed FLMM.



**Figure 7.** Fitted covariance surfaces via FPCA and proposed FLMM.

## 6. Discussion

In this paper, we propose a novel estimator of large functional data using the FLMM technique. In comparison with the FPCA, this novel estimator is much more efficient because all parameters can be automatically estimated. In comparison with the traditional estimator of FLMM, i.e., the REML criterion, our novel estimator adopts the LAML criterion, which enjoys a significantly lower computational complexity when the number of observations $T$ or the number of employed bases $p$ is large. In the simulation, our novel estimator of FLMM outperforms or performs equally as FPCA in all five criteria. In real data analysis, it is able to provide more reliable estimates than PFCA in terms of avoiding overfitting. Note that we only compare the novel estimator with the FPCA implemented by the PACE package, because the computing time of the traditional estimator of FLMM with REML is extremely large, and it is very easy to encounter numerical problems such as a degenerate Hessian matrix.

Another contribution of this paper is the asymptotic theory of FLMM. As far as we are concerned, our work is the first one to point out the convergence rates of mean function estimate, individual trajectory prediction, variance estimate, and covariance surface estimate. Theorems 1 and 2 show that the mean function estimate and individual trajectory prediction can reach the minimax lower bounds if the numbers of employed bases are chosen optimally. In particular, we point out that the number of basis functions should be adaptively chosen when estimating the mean function and individual trajectory, providing a novel guide about how to perform FLMM in practice. However, Theorems 3 and 4 illustrate that the convergence rates of variance estimate and covariance surface estimate cannot reach the minimax lower bounds because of the estimation errors of mean function estimate and individual trajectory prediction. These two findings indicate that the variance components cannot be precisely estimated if the mean function estimate and individual trajectory prediction are not consistent.

It should be pointed out that the proposed estimator can be simply extended to analyze unbalanced longitudinal data or the so-called sparse functional data [11]. Indeed, the FLMM estimator of unbalanced longitudinal data resembles the combination of penalized quasi-likelihood (PQL) and REML/LAML; a similar estimating procedure can be found in Breslow and Clayton [33]. Since in unbalanced longitudinal data, the number of observations for each individual is usually small, the FLMM estimator will not suffer from the theoretical and computational difficulties caused by the "dimensional curse". As mentioned before, Vonesh et al. [34] pointed out that the convergence rate of the mean parameter in GLMM is $O(T_{\min}^{-1/2})$. However, as far as we are concerned, the convergence rates of covariance matrix estimate and individual prediction are still unknown. As for FLMM, since the approximation bias of basis expansions is further involved, it is even unclear if the four estimates are consistent. Hence, it is worth future studying under what conditions the FLMM estimation is consistent for unbalanced longitudinal data or sparse functional data.

Discrete functional data analysis is an area that has received much less attention than continuous functional data analysis. However, discrete data are more common in longitudinal data analysis than continuous data, and hence the technical tools for analyzing discrete functional data are more urgently needed. Specifically, the model of discrete functional data is

$$\mathrm{E}(Y_i(t_j)) = g^{-1}(X_i(t_j) + \boldsymbol{W}_{ij}^\top \boldsymbol{\gamma}), \tag{50}$$

where $Y_i(t_j)$ is a random variable following the exponential family distribution [45], $g(\cdot)$ is the so-called link function, and $\boldsymbol{W}_{ij}$ is certain vector of covariates. We conjecture that the convergence rates of mean function estimate, individual trajectory prediction, variance estimate, and covariance surface estimate in this new model are the same as the ones indicated by Theorems 1–4 if the sampling locations $\{t_j\}$ are balanced for all individuals. On the other hand, such a problem has been studied by Hall et al. [46] using the FPCA

technique. It is also worth comparing the performances of FLMM and FPCA when dealing with such discrete functional data.

## Appendix A. Proofs of Theorems

*Appendix A.1. Lemmas*

We present some basic lemmas that facilitate the proofs of Theorems 1–4.

**Lemma A1.** *Suppose $X_1, \ldots, X_n$ are n independent sub-Gaussian variables with mean-zero and variance $\sigma_1^2, \ldots, \sigma_n^2$. Then*

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{D} \mathcal{N}(0, \sigma_x^2),$$

*where*

$$\sigma_x^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2.$$

**Proof of Lemma A1.** By using the second item of Proposition 2.5.2 of Vershynin [40],

$$[\mathrm{E}(|X_i|^p)]^{\frac{1}{p}} = O(\sqrt{p}) \tag{A1}$$

for all $p \geq 1$. Hence, for all fixed $\delta > 0$,

$$\lim_{n \to \infty} \frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathrm{E}(|X_i|^{2+2\delta}) \leq \frac{(K_0\sqrt{2+2\delta})^{2+2\delta}}{n^\delta} \to 0, \tag{A2}$$

where $K_0 > 0$ is a fixed number, which verifies Lyapunov's condition. □

**Lemma A2.** *Let $X_1, \ldots, X_n$ be n $(p \times 1)$ independent identically distributed random vector with entries $x_{i1}, \ldots, x_{ip}$ are sub-Gaussian with zero-mean. Besides, define the covariance matrix of $X_i$ as*

$$\boldsymbol{\Sigma} = E(\boldsymbol{X}_i \boldsymbol{X}_i^\top)$$

*and the related sample covariance matrix*

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^\top.$$

*Then for every positive integer n,*

$$E(||\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}||_2) \leq C\left(\frac{p}{n} + \sqrt{\frac{p}{n}}\right)||\boldsymbol{\Sigma}||_2,$$

*where C is a certain positive constant.*

This lemma is provided by Vershynin [40] (Theorem 4.7.1).

**Lemma A3.**

- *Weyl's lemma: Let* $\mathbf{S}, \mathbf{E}$ *be two Hermitian* $(p \times p)$ *matrices. Then, for each* $1 \leq i \leq p$,

$$\lambda_i(\mathbf{S}) + \lambda_{\max}(\mathbf{E}) \leq \lambda_i(\mathbf{S} + \mathbf{E}) \leq \lambda_i(\mathbf{S}) + \lambda_{\min}(\mathbf{E}),$$

- *Ostrowski's lemma: Let* $\mathbf{A}$ *be a Hermitian* $(n \times p)$ *matrix, and* $\mathbf{S}$ *be an* $(p \times p)$ *matrix. Then, for each* $1 \leq i \leq p$, *there exists an nonnegative real* $\theta_i$ *with* $\lambda_{\min}(\mathbf{A}\mathbf{A}^\top) \leq \theta_i \leq \lambda_{\max}(\mathbf{A}\mathbf{A}^\top)$ *such that*

$$\lambda_i(\mathbf{A}\mathbf{S}\mathbf{A}^\top) = \theta_i\lambda_i(\mathbf{S}),$$

*where* $\lambda_i(\mathbf{S})$ *means the ith eigenvalues of* $\mathbf{S}$.

This lemma can be found in Horn and Johnson [47] (Theorem 4.3.1 and Theorem 4.5.9).

**Lemma A4.** *Let* $\boldsymbol{\Gamma}$ *be an* $(p \times p)$ *symmetric matrix satisfying* $p\boldsymbol{\Gamma}$, *a well-conditioned covariance matrix. Then*

$$TC_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Psi}_q^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_q) \leq \lambda_{\max}(\boldsymbol{\Psi}_q^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_q) \leq C_0 T,$$

*where* $\mathbf{V} = \sigma^2\mathbf{I}_T + \boldsymbol{\Psi}_p\boldsymbol{\Gamma}\boldsymbol{\Psi}_p^\top$, $C_0 > 0$ *is a constant,* $\sigma^2 > 0$ *is an alternative constant,* $\boldsymbol{\Psi}_q$ *is an* $(T \times q)$ *matrix satisfying* $\boldsymbol{\Psi}_q^\top \boldsymbol{\Psi}_q/T \to \boldsymbol{\Sigma}_{\Psi_q}$, $\boldsymbol{\Psi}_p$ *is an* $(T \times p)$ *matrix satisfying* $\boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p/T \to \boldsymbol{\Sigma}_{\Psi_p}$, *and* $\boldsymbol{\Sigma}_{\Psi_q}$, $\boldsymbol{\Sigma}_{\Psi_p}$ *are two well-conditioned covariance matrices.*

**Proof of Lemma A4.** By the Ostrowski's lemma,

$$0 \leq \lambda_{\min}(\boldsymbol{\Psi}_p\boldsymbol{\Gamma}\boldsymbol{\Psi}_p^\top), \quad \lambda_{\max}(\boldsymbol{\Psi}_p\boldsymbol{\Gamma}\boldsymbol{\Psi}_p^\top) \leq \lambda_{\max}(T\boldsymbol{\Sigma}_{\Psi_p})\lambda_{\max}(\boldsymbol{\Gamma}) \leq \frac{c_0 T}{T} = c_0, \tag{A3}$$

for certain constant $c_0$, because the minimum eigenvalue of $\boldsymbol{\Psi}_p\boldsymbol{\Psi}_p^\top$ is 0. By using Weyl's lemma,

$$\sigma^2 \leq \lambda_{\min}(\mathbf{V}) \leq \lambda_{\min}(\mathbf{V}) \leq \sigma^2 + c_0. \tag{A4}$$

Hence,

$$\frac{T\sigma^{-2}\lambda_{\max}(\boldsymbol{\Sigma}_{\Psi_q})}{\sigma^2 + c_0} \leq \lambda_{\min}(T\boldsymbol{\Sigma}_{\Psi_q})\lambda_{\min}(\mathbf{V}^{-1}) \leq \lambda_{\min}(\boldsymbol{\Psi}_q^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_q)$$
$$\lambda_{\max}(\boldsymbol{\Psi}_q^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_q) \leq \lambda_{\max}(T\boldsymbol{\Sigma}_{\Psi_q})\lambda_{\max}(\mathbf{V}^{-1}) = T\sigma^{-2}\lambda_{\max}(\boldsymbol{\Sigma}_{\Psi_q}). \tag{A5}$$

That is,

$$TC_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Psi}_q^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_q) \leq \lambda_{\max}(\boldsymbol{\Psi}_q^\top \mathbf{V}^{-1}\boldsymbol{\Psi}_q) \leq C_0 T, \tag{A6}$$

for certain positive constant $C_0$. $\quad\square$

*Appendix A.2. Proofs*

**Proof of Theorem 1.** The score function of $\hat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
\mathbf{0} &= \frac{\partial \frac{1}{N}\{\sum_{i=1}^{N}(\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\boldsymbol{\beta})^\top \hat{\boldsymbol{\Sigma}}_y^{-1}(\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\boldsymbol{\beta}) + \gamma\boldsymbol{\beta}^\top\mathbf{Q}\boldsymbol{\beta}\}}{\partial\boldsymbol{\beta}}\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \\
&= -\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Y}_i - \boldsymbol{\Psi}_q\hat{\boldsymbol{\beta}}\right) \\
&= -\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Y}_i - \boldsymbol{\mu}\right) - \boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}(\boldsymbol{\mu} - \boldsymbol{\Psi}_q\boldsymbol{\beta}) \\
&\quad - \boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Psi}_q(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \frac{\gamma}{N}\mathbf{Q}\boldsymbol{\beta} + \frac{\gamma}{N}\mathbf{Q}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad\qquad\text{(A7)} \\
&= I_1 + I_2 + I_3 + I_4 + I_5. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\text{(A8)}
\end{aligned}
$$

As for $I_1$,

$$
\begin{aligned}
I_1 &= -\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{Y}_i - \boldsymbol{\mu}\right) = -\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Sigma}_y^{\frac{1}{2}}\boldsymbol{\Sigma}_y^{-\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{X}_i + \boldsymbol{\varepsilon}_i)\right) \\
&= -\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Sigma}_y^{\frac{1}{2}}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{e}_i\right),
\end{aligned}
\qquad\text{(A9)}
$$

where $\boldsymbol{e} = (e_{i1},\ldots,e_{iT})^\top$ with $\mathrm{E}(e_{ij}) = 0$, var$(e_{ij})$=1, and $\mathrm{E}(e_{ij}e_{ik}) = 0$ for $j \neq k$. Here,

$$
\mathrm{E}(I_1) = \mathbf{0}, \quad \mathrm{var}(I_1) = \frac{1}{N}\times\mathrm{trace}(\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Sigma}_y^{\frac{1}{2}}\mathbf{I}_T\boldsymbol{\Sigma}_y^{\frac{1}{2}}\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Psi}_q) \qquad\text{(A10)}
$$

which guarantees

$$
\begin{aligned}
\mathrm{E}(||I_1||_2^2) &\leq \frac{1}{N}\times\mathrm{trace}(\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Psi}_q)\times\lambda_{\max}(\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Sigma}_y) \\
&= O\left(\frac{Tq}{N}\right).
\end{aligned}
\qquad\text{(A11)}
$$

As for $I_2$, by Assumption 4,

$$
\begin{aligned}
||I_2||_2^2 &\leq \lambda_{\max}(\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y^{-1}\boldsymbol{\Psi}_q)\times\lambda_{\min}^{-1}(\hat{\boldsymbol{\Sigma}}_y)\times||\boldsymbol{\mu} - \boldsymbol{\Psi}_q\boldsymbol{\beta}||_2^2 \\
&= O(T)\times O(1)\times O(Tq^{-2s}) = O(T^2q^{-2s}).
\end{aligned}
\qquad\text{(A12)}
$$

As for $I_3 + I_5$,

$$
I_3 + I_5 \geq \lambda_{\min}\left(\boldsymbol{\Psi}_q^\top\hat{\boldsymbol{\Sigma}}_y\boldsymbol{\Psi}_q + \frac{\gamma}{N}\mathbf{Q}\right)\times||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2 \geq O(T)\times||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2.
$$

As for $I_4$,

$$
||I_4||_2^2 \leq \frac{1}{N^2}\times\lambda_{\max}^2(\gamma\mathbf{Q})\times||\boldsymbol{\beta}||_2^2 = O\left(\frac{1}{N^2}\right) \qquad\text{(A13)}
$$

As $q \leq T$ when estimating the mean function,

$$
||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 \leq O_P\left(\frac{1}{T^2}(||I_1||_2^2 + ||I_3||_2^2)\right) = O_P\left(\frac{q}{NT} + \frac{1}{q^{2s}}\right). \qquad\text{(A14)}
$$

Hence, by letting $q = O(T)$,

$$||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 = O_P\left(\frac{1}{N} + T^{-2s}\right), \tag{A15}$$

which is consistent with the minimax lower bounder provided by Cai and Yuan [29]. Therefore, Theorem 1 is proved. □

**Proof of Theorem 2.** The score function of $\hat{\boldsymbol{c}}_i$ is

$$
\begin{aligned}
\boldsymbol{0} &= \left.\frac{\partial\{||\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p\boldsymbol{c}_i||_2^2 + \boldsymbol{c}_i^\top\hat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{c}_i\}}{\partial\boldsymbol{c}_i}\right|_{\boldsymbol{c}_i=\hat{\boldsymbol{c}}_i} = -\boldsymbol{\Psi}_p^\top(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p\hat{\boldsymbol{c}}_i) + \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{c}}_i \\
&= -\boldsymbol{\Psi}_p^\top(\boldsymbol{Y}_i - \boldsymbol{X}_i) - \boldsymbol{\Psi}_p^\top(\boldsymbol{X}_i - \boldsymbol{\Psi}_p\boldsymbol{b}_i) \\
&\quad - \boldsymbol{\Psi}_q^\top\boldsymbol{\Psi}_q(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \boldsymbol{\Psi}_p^\top\boldsymbol{\Psi}_p(\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i) + \hat{\boldsymbol{\Lambda}}\boldsymbol{c}_i + \hat{\boldsymbol{\Lambda}}(\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i) \\
&= J_1 + J_2 + J_3 + J_4 + J_5 + J_6,
\end{aligned}
\tag{A16}
$$

where $\hat{\boldsymbol{\Lambda}} = \hat{\sigma}^2\hat{\boldsymbol{\Gamma}}^{-1}$. As for $J_1$, it is easy to see

$$\mathrm{E}(J_1) = \boldsymbol{0}, \quad \mathrm{var}(J_1) = \sigma^2\mathrm{trace}(\boldsymbol{\Psi}_p^\top\boldsymbol{I}_T\boldsymbol{\Psi}_p) = O(pT), \tag{A17}$$

which guarantees

$$\mathrm{E}(||J_1||_2^2) = O(pT). \tag{A18}$$

As for $J_2$, by Assumption 4,

$$\mathrm{E}(||J_2||_2^2) \leq \lambda_{\max}(\boldsymbol{\Psi}_p^\top\boldsymbol{\Psi}_p)\mathrm{E}(||\boldsymbol{X}_i - \boldsymbol{\Psi}_p\boldsymbol{b}_i||_2^2) = O_P(T^2p^{-2s}). \tag{A19}$$

As for $J_3$,

$$||J_3||_2^2 \leq \lambda_{\max}(\boldsymbol{\Psi}_q^\top\boldsymbol{\Psi}_q)||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 = O_P\left(\frac{T}{N} + T^{1-2s}\right). \tag{A20}$$

As for $J_5$,

$$||J_5||_2^2 \leq \lambda_{\max}(\hat{\boldsymbol{\Lambda}})\boldsymbol{c}_i^\top\hat{\boldsymbol{\Lambda}}\boldsymbol{c}_i. \tag{A21}$$

Using the consequence that $\hat{\boldsymbol{\Lambda}} \to \sigma^2\boldsymbol{\Gamma}^{-1}$,

$$\boldsymbol{c}_i^\top\hat{\boldsymbol{\Lambda}}\boldsymbol{c}_i \sim \sigma^2\chi_p^2. \tag{A22}$$

Hence,

$$||J_5||_2^2 = O_P(T) \times O_P(p) = O_P(Tp). \tag{A23}$$

As for $J_4 + J_6$,

$$
\begin{aligned}
||J_4 + J_6||_2^2 &\geq (\lambda_{\min}(\boldsymbol{\Psi}_p^\top\boldsymbol{\Psi}_p) + \lambda_{\max}^{-1}(\hat{\boldsymbol{\Lambda}}))^2||\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i||_2^2 \geq O_P(T^2)||\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i||_2^2 \\
&\geq O_P(T^2)||\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i||_2^2.
\end{aligned}
\tag{A24}
$$

As a result,

$$||\hat{\boldsymbol{c}}_i - \boldsymbol{c}_i||_2^2 \leq O_P\left(\frac{p}{T} + \frac{1}{p^{2s}} + \frac{1}{NT} + \frac{1}{T^{2s+1}} + \frac{p}{T}\right) \tag{A25}$$

By letting $p = O(T^{1/(2s+1)})$,

$$||\hat{c}_i - c_i||_2^2 \leq O_P(T^{-\frac{2s}{2s+1}}).$$ (A26)

Thus, Theorem 2 is proved. □

**Proof of Theorem 3.** The score function of $\hat{\sigma}^2$ is

$$
\begin{aligned}
0 = \frac{\partial \mathcal{L}(\sigma^2, \boldsymbol{\Gamma})}{NT \partial \sigma^{-2}}\bigg|_{\sigma^2 = \hat{\sigma}^2} &= -(\hat{\sigma}^2 - \sigma^2) + \left(\frac{1}{NT}\sum_{i=1}^{N}||\boldsymbol{Y}_i - \boldsymbol{\mu} - \boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2 - \sigma^2\right) \\
&+ \frac{1}{NT}\sum_{i=1}^{N}\left(||\boldsymbol{Y}_i - \boldsymbol{\mu} - \boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2 - ||\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2\right) \\
&+ \frac{1}{NT}\sum_{i=1}^{N}\left(||\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p \hat{\boldsymbol{c}}_i||_2^2 - ||\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2\right) + \frac{\text{trace}(\boldsymbol{\Delta}(\sigma^2, \boldsymbol{\Gamma})\boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p)}{NT} \\
&= K_1 + K_2 + K_3 + K_4 + K_5.
\end{aligned}
$$ (A27)

As for $K_2$, it is easy to see

$$K_2 = \frac{1}{NT}\sum_{i=1}^{N}||\boldsymbol{Y}_i - \boldsymbol{\mu} - \boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2 - \sigma^2 = (\hat{\sigma}^2_{\text{oracle}} - \sigma^2),$$ (A28)

where

$$\hat{\sigma}^2_{\text{oracle}} = \frac{1}{NT}\sum_{i=1}^{N}||\boldsymbol{Y}_i - \boldsymbol{\mu} - \boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2.$$ (A29)

It is well known that $||\hat{\sigma}^2_{\text{oracle}} - \sigma^2||^2 = O_P((NT)^{-1})$ hence $K_2^2 = O_P((NT)^{-1})$. As for $K_3$,

$$K_3 = \frac{1}{NT}\sum_{i=1}^{N}(\boldsymbol{\mu} - \hat{\boldsymbol{\mu}})^\top(2\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} - 2\boldsymbol{\Psi}_p \boldsymbol{c}_i).$$ (A30)

Therefore,

$$
\begin{aligned}
K_3^2 &\leq \left(\frac{1}{NT}\sum_{i=1}^{N}||\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}||_2^2\right)\left(\frac{1}{NT}\sum_{i=1}^{N}||2\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i - 2\boldsymbol{\Psi}_p \boldsymbol{c}_i||_2^2\right) \\
&= O_P\left(\frac{1}{N} + \frac{1}{T^{2s}}\right).
\end{aligned}
$$ (A31)

As for $K_4$,

$$K_4 = \frac{1}{NT}\sum_{i=1}^{N}(\boldsymbol{c}_i - \hat{\boldsymbol{c}}_i)^\top \boldsymbol{\Psi}_p^\top(2\boldsymbol{Y}_i - 2\hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p \boldsymbol{c}_i - \boldsymbol{\Psi}_p \hat{\boldsymbol{c}}_i).$$ (A32)

Therefore,

$$
\begin{aligned}
K_4^2 &\leq \frac{1}{N^2T^2}\left(\sum_{i=1}^{N}||2\boldsymbol{Y}_i - 2\hat{\boldsymbol{\mu}} - \boldsymbol{\Psi}_p \boldsymbol{c}_i - \boldsymbol{\Psi}_p \hat{\boldsymbol{c}}_i||_2^2\right)\left(\sum_{i=1}^{N}\lambda_{\max}(T\boldsymbol{\Sigma}_{\Psi_p})||\boldsymbol{c}_i - \hat{\boldsymbol{c}}_i||_2^2\right) \\
&= O_P\left(\frac{p}{T} + \frac{1}{p^{2s}}\right).
\end{aligned}
$$ (A33)

As for $K_5$,

$$K_5^2 \leq \frac{1}{N^2T^2} \times \lambda_{\max}(\boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p) \times \lambda_{\min}^{-1}(\boldsymbol{\Psi}_p^\top \boldsymbol{\Psi}_p) = O_P\left(\frac{1}{N^2T^2}\right).$$ (A34)

As a result,

$$(\hat{\sigma}^2 - \sigma^2)^2 = O(K_4) = O_P\left(\frac{p}{T} + \frac{1}{p^{2s}}\right). \tag{A35}$$

Hence, letting $p = O(T^{1/(2s+1)})$,

$$(\hat{\sigma}^2 - \sigma^2)^2 = O_P\left(T^{-\frac{2s}{2s+1}}\right). \tag{A36}$$

Thus, Theorem 3 is proved. □

**Proof of Theorem 4.** The score function of $\hat{\Gamma}$ is

$$\begin{aligned}
0 &= \frac{\frac{1}{N}\partial\mathcal{L}(\sigma^2, \Gamma)}{\partial\Gamma^{-1}}\bigg|_{\Gamma^{-1}=\hat{\Gamma}^{-1}} = -\hat{\Gamma} + \frac{1}{N}\sum_{i=1}^N \hat{c}_i\hat{c}_i^\top + \frac{1}{N}\Delta(\sigma^2, \Gamma) \\
&= -(\hat{\Gamma} - \Gamma) + \left(\frac{1}{N}\sum_{i=1}^N c_i c_i^\top - \Gamma\right) + \left(\frac{1}{N}\sum_{i=1}^N \hat{c}_i\hat{c}_i^\top - \frac{1}{N}\sum_{i=1}^N c_i c_i^\top\right) \\
&\quad + \frac{1}{N}\Delta(\sigma^2, \Gamma) = L_1 + L_2 + L_3 + L_4.
\end{aligned} \tag{A37}$$

As for $L_2$, by using Vershynin [40] (Theorem 4.7.1),

$$||L_2||_2^2 = \left\|\frac{1}{N}\sum_{i=1}^N c_i c_i^\top - \Gamma\right\|_2^2 = O_P\left(\frac{p}{N}\right) \times ||\Gamma||_2^2 = O_P\left(\frac{p}{NT}\right). \tag{A38}$$

As for $L_3$,

$$L_3 = \frac{1}{N}\sum_{i=1}^N \hat{c}_i(\hat{c}_i - c_i)^\top + \frac{1}{N}\sum_{i=1}^N (\hat{c}_i - c_i)c_i^\top = L_{31} + L_{32}. \tag{A39}$$

For $L_{31}$,

$$||L_{31}||_2^2 \leq \lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^N \hat{c}_i\hat{c}_i^\top\right)\left(\frac{1}{N}\sum_{i=1}^N ||\hat{c}_i - c_i||^2\right) \leq O_P\left(\frac{p}{T^2} + \frac{1}{p^{2s}T}\right). \tag{A40}$$

As for $L_{32}$,

$$||L_{32}||_2^2 \leq \lambda_{\max}\left(\frac{1}{N}\sum_{i=1}^N c_i c_i^\top\right)\left(\frac{1}{N}\sum_{i=1}^N ||\hat{c}_i - c_i||^2\right) \leq O_P\left(\frac{p}{T^2} + \frac{1}{p^{2s}T}\right). \tag{A41}$$

As for $L_4$,

$$||L_4||_2^2 \leq N^{-2}\lambda_{\min}^{-1}(\Psi_p^\top\Psi_p) = O_P(T^{-1}N^{-2}). \tag{A42}$$

Note that $L_2$ dominates $L_3$ and $L_4$. Therefore,

$$||\hat{\Gamma} - \Gamma||_2^2 = O_P\left(\frac{p}{NT} + \frac{p}{T^2} + \frac{1}{p^{2s}T}\right). \tag{A43}$$

Hence, letting $p = O(T^{1/(2s+1)})$,

$$||\Gamma^{-\frac{1}{2}}\hat{\Gamma}\Gamma^{-\frac{1}{2}} - I_p||_2^2 = O_P\left(N^{-1}T^{\frac{1}{2s+1}} + T^{-\frac{2s}{2s+1}}\right) \tag{A44}$$

Thus, Theorem 4 is proved. □

## Appendix B. Estimators of FPCA and FLMM

*Appendix B.1. Estimator of FPCA*

In implementation, FPCA will (1) estimate the mean function $\mu(t)$; (2) estimate the two-dimensional covariance function $C(s,t)$; (3) yield the first $K$ eigenfunctions and eigenvalues $\{\phi_k(t)\}$ and $\{\lambda_k\}$ through PCA; and (4) predict the corresponding coordinates $\{\xi_{ik}\}$ by using Gaussian conditional expectation. Without loss of generality, we illustrate these four steps with the procedure employed by Li and Hsing [10]. Specifically, Li and Hsing employed the local linear modeling [30] to estimate $\mu(t)$:

$$(\hat{a}_0, \hat{a}_1) = \arg\min_{a_0,a_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{j=1}^{T_i} \left[ Y_{ij} - a_0 - a_1(t_{ij} - t) \right]^2 K_{h_u}(t_{ij} - t) \right\}, \tag{A45}$$

where $K_{h_u}(t)$ is a kernel function with bandwidth $h_u$. The estimate $\hat{u}(t)$ is just $\hat{a}_0$. Subsequently, they estimate $C(s,t)$ by

$$(\hat{b}_0, \hat{b}_1, \hat{b}_2) = \arg\min_{b_0,b_1,b_2} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{j=1}^{T_i} \sum_{k=1}^{T_i} \left[ Y_{ij}Y_{ik} - b_0 - b_1(t_{ij} - s) - b_2(t_{ik} - t) \right]^2 \right. $$
$$\left. \times K_{h_C}(t_{ij} - s) K_{h_C}(t_{ik} - t) \right\}, \tag{A46}$$

where $K_{h_C}(t)$ is a kernel function with bandwidth $h_C$. The estimate $\hat{C}(s,t)$ is given by $\hat{b}_0 - \hat{\mu}(t)\hat{\mu}(s)$. Next, the variance of random error is estimated using a two-step produce. In the former step, Li and Hsing will estimate the variance function

$$(\hat{c}_0, \hat{c}_1) = \arg\min_{c_0,c_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{j=1}^{T_i} \left[ Y_{ij}^2 - c_0 - c_1(t_{ij} - t) \right]^2 K_{h_v}(t_{ij} - t) \right\}, \tag{A47}$$

where $K_{h_v}(t)$ is a kernel function with bandwidth $h_v$. The estimate $\hat{V}(t)$ is given by $\hat{c}_0$. In the latter step, $\hat{\sigma}$ is yielded by

$$\hat{\sigma}^2 = \int_0^1 \left( \hat{V}(t) - \hat{C}(t,t) - \hat{\mu}(t)^2 \right) dt. \tag{A48}$$

Furthermore, $\{\hat{\phi}_k(t)\}$ and $\{\hat{\lambda}_k\}$ are yielded by

$$\hat{C}(s,t) = \sum_{k=1}^{K} \hat{\lambda}_k \hat{\phi}_k(s) \hat{\phi}_k(t)^\top, \tag{A49}$$

where $K$ is a pre-given cutoff that can be selected by information criterion such as Bayesian information criterion (BIC) [48]. Finally, the random coordinate $\hat{\xi}_{ik}$ is predicted by

$$\hat{\xi}_{ik} = \hat{\mathbb{E}}(\xi_{ik}|\boldsymbol{Y}_i) = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^\top \hat{\boldsymbol{\Sigma}}_{y_i}^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i), \tag{A50}$$

where $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{iT_i})^\top$, $\hat{\boldsymbol{\phi}}_{ik} = (\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{iT_i}))^\top$, $\hat{\boldsymbol{\mu}}_i = (\hat{\mu}(t_{i1}), \ldots, \hat{\mu}(t_{iT_i}))^\top$, $\hat{\boldsymbol{\Sigma}}_i = \hat{\boldsymbol{C}}_i + \hat{\sigma}^2 \boldsymbol{I}_{T_i}$, and the $(k,h)$th element of $\hat{\boldsymbol{C}}_i$ is $\hat{C}(t_{ik}, t_{ih})$.

*Appendix B.2. Estimator of FLMM*

In the implementation, FLMM will (1) estimate the mean function $\mu(t)$; (2) estimate the covariance matrix $\boldsymbol{\Gamma}$ by REML, and (3) horizontally yield the covariance function $C(s,t)$ and predict the random coordinates $\{b_{jk}\}$ by using the estimate of $\boldsymbol{\Gamma}$. Without loss of generality,

we illustrate these three steps with the procedure employed by Shi et al. [20]. Specifically, Shi et al. first estimated $\mu(t)$ by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{Y}_i - \boldsymbol{\Psi}_i \boldsymbol{\beta})^{\top} \boldsymbol{\Sigma}_{y_i}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\Psi}_i \boldsymbol{\beta}) \right\}, \tag{A51}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\top}$, $\boldsymbol{\Psi}_i = (\Psi_1(\boldsymbol{t}_i), \ldots, \Psi_p(\boldsymbol{t}_i))$ is an $(T_i \times p)$ matrix, $\Psi_k(\boldsymbol{t}_i) = (\Psi_k(t_{i1}), \ldots, \Psi_k(t_{iT_i}))^{\top}$, $\text{cov}(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}_{y_i}$ is approximated by

$$\boldsymbol{\Sigma}_{y_i} \approx \sigma^2 \mathbf{I}_{T_i} + \boldsymbol{\Psi}_i \boldsymbol{\Gamma} \boldsymbol{\Psi}_i^{\top}, \tag{A52}$$

and $\hat{\boldsymbol{\Sigma}}_{y_i}$ is an estimate of $\boldsymbol{\Sigma}_{y_i}$. Then $\hat{\mu}(t) = \sum_{k=1}^{p} \Psi_k(t)\hat{\beta}_k$. Next, Shi et al. employed the technique of LMM [8] such as REML [7] to estimate $\sigma^2$ and $\boldsymbol{\Gamma}$:

$$\begin{aligned} (\hat{\sigma}^2, \hat{\boldsymbol{\Gamma}}) = \arg\min_{\sigma^2, \boldsymbol{\Gamma}} \bigg\{ \sum_{i=1}^{N} \bigg( & (\boldsymbol{Y}_i - \boldsymbol{\Psi}_i \hat{\boldsymbol{\beta}})^{\top} \boldsymbol{\Sigma}_{y_i}^{-1} (\boldsymbol{Y}_i - \boldsymbol{\Psi}_i \hat{\boldsymbol{\beta}}) \\ & + \log \det(\boldsymbol{\Sigma}_{y_i}) + \log \det(\boldsymbol{\Psi}_i^{\top} \boldsymbol{\Sigma}_{y_i}^{-1} \boldsymbol{\Psi}_i) \bigg) \bigg\}. \end{aligned} \tag{A53}$$

FLMM will iteratively implement the above two steps until the stable estimates of $\hat{\beta}_1, \ldots, \hat{\beta}_p$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\Gamma}}$ is obtained. Here, the estimate of covariance function $C(s, t)$ is

$$\hat{C}(s, t) = \sum_{k=1}^{p} \sum_{h=1}^{p} \hat{\Gamma}_{kh} \Psi_k(s) \Psi_h(t). \tag{A54}$$

As for the random coordinate vector $\boldsymbol{b}_i = (b_{j1}, \ldots, b_{jp})^{\top}$, its prediction is given by

$$\hat{\boldsymbol{b}}_i = \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\Gamma}} \boldsymbol{\Psi}_i^{\top} \hat{\boldsymbol{\Sigma}}_{y_i}^{-1} (\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i). \tag{A55}$$

The individual trajectory is predicted by $\hat{X}_i(t) = \sum_{k=1}^{p} \Psi_k(t)\hat{b}_{ik}$.

## References

1. Ramsay, J.; Silverman, B. *Functional Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2005.
2. Králík, M.; Klíma, O.; Ǒuta, M.; Malina, R.M.; Kozieł, S.; Polcerová, L.; Škultétyová, A.; Španěl, M.; Kukla, L.; Zemčík, P. Estimating Growth in Height from Limited Longitudinal Growth Data Using Full-Curves Training Dataset: A Comparison of Two Procedures of Curve Optimization—Functional Principal Component Analysis and SITAR. *Children* **2021**, *8*, 934. [CrossRef] [PubMed]
3. Ullah, S.; Finch, C. Applications of functional data analysis: A systematic review. *BMC Med Res. Methodol.* **2013**, *13*, 43. [CrossRef] [PubMed]
4. Cai, T.; Yuan, M. *Nonparametric Covariance Function Estimation for Functional and Longitudinal Data*; University of Pennsylvania and Georgia Inistitute of Technology: Philadelphia, PA, USA, 2010.
5. Diggle, P.; Diggle, P.; Heagerty, P.; Liang, K.; Zeger, S. *Analysis of Longitudinal Data*; Oxford University Press: Oxford, UK, 2002.
6. Wood, S.; Wood, M. Package 'mgcv'. *R Package Version* **2015**, *1*, 729.
7. Patterson, H.; Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **1971**, *58*, 545–554. [CrossRef]
8. Laird, N.; Ware, J. Random-effects models for longitudinal data. *Biometrics* **1982**, *1*, 963–974. [CrossRef]
9. Hall, P.; Müller, H.; Wang, J. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Stat.* **2006**, *34*, 1493–1517. [CrossRef]
10. Li, Y.; Hsing, T. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Ann. Stat.* **2010**, *38*, 3321–3351. [CrossRef]
11. Yao, F.; Müller, H.; Wang, J. Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.* **2005**, *100*, 577–590. [CrossRef]
12. Paul, D.; Peng, J. Consistency of restricted maximum likelihood estimators of principal components. *Ann. Stat.* **2009**, *37*, 1229–1271. [CrossRef]
13. Peng, J.; Paul, D. A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *J. Comput. Graph. Stat.* **2009**, *18*, 995–1015. [CrossRef]

14. Bunea, F.; Ivanescu, A.; Wegkamp, M. Adaptive inference for the mean of a Gaussian process in functional data. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2011**, *73*, 531–558. [CrossRef]
15. Rice, J.; Silverman, B. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1991**, *53*, 233–243. [CrossRef]
16. Yu, F.; Liu, L.; Yu, N.; Ji, L.; Qiu, D. A method of L1-norm principal component analysis for functional data. *Symmetry* **2020**, *12*, 182. [CrossRef]
17. James, G.; Hastie, T.; Sugar, C. Principal component models for sparse functional data. *Biometrika* **2000**, *87*, 587–602. [CrossRef]
18. James, G.; Sugar, C. Clustering for sparsely sampled functional data. *J. Am. Stat. Assoc.* **2003**, *98*, 397–408. [CrossRef]
19. Rice, J.; Wu, C. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **2001**, *57*, 253–259. [CrossRef]
20. Shi, M.; Weiss, R.; Taylor, J. An analysis of paediatric CD4 counts for acquired immune deficiency syndrome using flexible random curves. *J. R. Stat. Soc. Ser. C Appl. Stat.* **1996**, *45*, 151–163. [CrossRef]
21. Antoniadis, A.; Sapatinas, T. Estimation and inference in functional mixed-effects models. *Comput. Stat. Data Anal.* **2007**, *51*, 4793–4813. [CrossRef]
22. Morris, J.; Carroll, R. Wavelet-based functional mixed models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2006**, *68*, 179–199. [CrossRef]
23. Zhu, H.; Brown, P.; Morris, J. Robust, adaptive functional regression in functional mixed model framework. *J. Am. Stat. Assoc.* **2011**, *106*, 1167–1179. [CrossRef]
24. Qiu, P.; Zou, C.; Wang, Z. Nonparametric profile monitoring by mixed effects modeling. *Technometrics* **2010**, *52*, 265–277. [CrossRef]
25. Wu, H.; Zhang, J. Local polynomial mixed-effects models for longitudinal data. *J. Am. Stat. Assoc.* **2002**, *97*, 883–897. [CrossRef]
26. Wu, H.; Zhang, J. *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-Effects Modeling Approaches*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
27. Rice, J. Functional and longitudinal data analysis: Perspectives on smoothing. *Stat. Sinica* **2004**, *1*, 631–647.
28. Wang, J.; Chiou, J.; Müller, H. Functional data analysis. *Annu. Rev. Stat. Appl.* **2016**, *3*, 257–295. [CrossRef]
29. Cai, T.; Yuan, M. Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *Ann. Stat.* **2011**, *39*, 2330–2355. [CrossRef]
30. Fan, J.; Gijbels, I. *Local Polynomial Modelling and Its Applications*; Routledge: London, UK, 2018.
31. Eilers, P.; Marx, B. Flexible smoothing with B-splines and penalties. *Stat. Sci.* **1996**, *11*, 89–102. [CrossRef]
32. Acal, C.; Aguilera, A.; Escabias, M. New modeling approaches based on varimax rotation of functional principal components. *Mathematics* **2020**, *8*, 2085. [CrossRef]
33. Breslow, N.; Clayton, D. Approximate inference in generalized linear mixed models. *J. Am. Stat. Assoc.* **1993**, *88*, 9–25.
34. Vonesh, E.; Wang, H.; Nie, L.; Majumdar, D. Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *J. Am. Stat. Assoc.* **2002**, *97*, 271–283. [CrossRef]
35. Breslow, N.; Lin, X. Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **1995**, *82*, 81–91. [CrossRef]
36. Lee, Y.; Nelder, J. Hierarchical generalized linear models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **1996**, *58*, 619–656. [CrossRef]
37. Lin, X.; Breslow, N. Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Am. Stat. Assoc.* **1996**, *91*, 1007–1016. [CrossRef]
38. Ruppert, D.; Wand, M.; Carroll, R. *Semiparametric Regression*; Cambridge University Press: Cambridge, UK, 2003.
39. Wood, S.; Pya, N.; Säfken, B. Smoothing parameter and model selection for general smooth models. *J. Am. Stat. Assoc.* **2016**, *111*, 1548–1563. [CrossRef]
40. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*; Cambridge University Press: Cambridge, UK, 2018.
41. Bickel, P.; Levina, E. Regularized estimation of large covariance matrices. *Ann. Stat.* **2008**, *36*, 199–227. [CrossRef]
42. Wainwright, M. *High-Dimensional Statistics: A Non-asymptotic Viewpoint*; Cambridge University Press: Cambridge, UK, 2019.
43. Tsybakov, A.B. *Introduction to Nonparametric Estimation*; Springer: Berlin/Heidelberg, Germany, 2009.
44. Wood, S. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2011**, *73*, 3–36. [CrossRef]
45. Nelder, J.; Wedderburn, R. Generalized linear models. *J. R. Stat. Soc. Ser. A (Gen.)* **1972**, *135*, 370–384. [CrossRef]
46. Hall, P.; Müller, H.; Yao, F. Modelling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2008**, *70*, 703–723. [CrossRef]
47. Horn, R.; Johnson, C. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012.
48. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *1*, 461–464. [CrossRef]