Computational Statistics and Data Analysis 97 (2016) 1–15

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Mixture of functional linear models and its application to CO_2-GDP functional data

Shaoli Wang [a], Mian Huang [a], Xing Wu [a], Weixin Yao [b,*]

[a] School of Statistics and Management, and Shanghai Key Laboratory of Financial Information Technology, Shanghai University of Finance and Economics, Shanghai, 200433, PR China

[b] Department of Statistics, University of California, Riverside, CA, 92521, USA

ARTICLE INFO

Article history:
Received 29 June 2014
Received in revised form 11 November 2015
Accepted 15 November 2015
Available online 28 November 2015

Keywords:
Mixtures of functional linear regressions
Identifiability
EM-type algorithm
Kernel regression
Functional principal component analysis
Conditional bootstrap
Hypothesis test

ABSTRACT

Functional linear models are important tools for studying the relationship between functional response and covariates. However, if subjects come from an inhomogeneous population that demonstrates different linear relationship between the response and covariates among different subpopulations/clusters, a single functional linear model is no longer adequate for the data. A new class of mixtures of functional linear models for the analysis of heterogeneous functional data is introduced. Identifiability is established for the proposed class of mixture models under mild conditions. The proposed estimation procedures combine the ideas of local kernel regression, functional principal component analysis and EM algorithm. A generalized likelihood ratio test based on a conditional bootstrap is given as to whether the regression coefficient functions are constant. A Monte Carlo simulation study is conducted to examine the finite sample performance of the new methodology. Finally, the analysis of CO_2-GDP data reveals the dynamic patterns of relationship between CO_2 and GDP among different countries.

1. Introduction

The observations of functional data are functions defined over some set T. The last two decades have seen tremendous advances in functional data analysis. Ramsay and Silverman (2002, 2005) offer an excellent introduction to functional data analysis. Ferraty and Vieu (2006) give a detailed account of nonparametric methods for functional data. Horváth and Kokoszka (2012) focus on statistical inference for functional data, particularly on hypothesis tests in various functional data settings. Ferraty and Romain (2011) contain a comprehensive and up-to-date review on a broad range of topics in functional data analysis. Bongiorno et al. (2014) bring in the recent advances in functional data analysis and related areas. Bosq (2000) and Bosq and Blanke (2007) lay the mathematical foundations for functional data analysis.

In regression analysis for functional data, either the response or covariates, or both can be functions. Functional linear models (FLMs) (Ramsay and Silverman, 2005) are useful for modeling the linear relationship between a scalar response and functional covariates. However, FLMs fail for nonlinear regression functions, and nonparametric techniques have been employed in the literature. Ferraty et al. (2013) introduce the functional projection pursuit regression. Chen et al. (2011) study single and multiple index functional regression models with nonparametric link functions. Kudraszow and Vieu (2013) propose a kNN generalized regression estimator for the regression function and prove its uniform consistency. When both

* Corresponding author.
E-mail address: weixin.yao@ucr.edu (W. Yao).

http://dx.doi.org/10.1016/j.csda.2015.11.008
0167-9473/

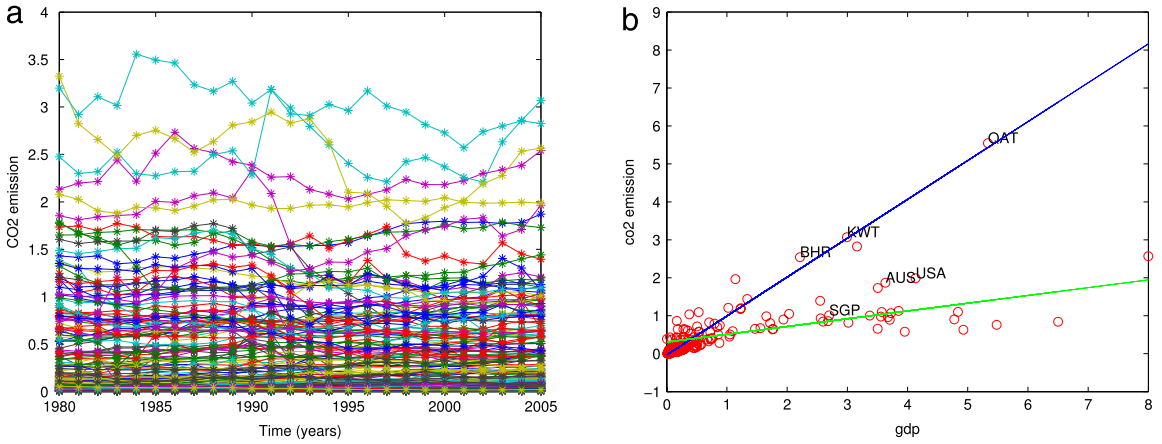

Fig. 1. (a) Observed CO_2 emission trajectories of 175 countries; (b) Scatter plot of CO_2-GDP sub-dataset of year 2005 with two fitted regression lines. y: CO_2 emission per capita; x: GDP per capita.

multivariate and functional covariates are present, semi-functional partial linear regression models can be used (Aneiros-Pérez and Vieu, 2006).

When both the response and covariates are functions, we can use a specific class of FLMs, called concurrent FLMs:

$$y(t) = \boldsymbol{X}(t)^T \boldsymbol{\beta}(t) + \varepsilon(t), \quad t \in T, \tag{1.1}$$

where $y(t)$ is a univariate response process, $\boldsymbol{X}(t)$ a p-dimensional covariate process, $\boldsymbol{\beta}(t)$ the unknown smooth regression coefficient function, and $\varepsilon(t)$ a zero mean error process that is independent of $\boldsymbol{X}(t)$. The time set T is typically assumed to be a closed bounded interval. In the concurrent model (1.1), the functional response $y(t)$ depends on the functional covariates $\boldsymbol{X}(t)$ in a point-wise manner. This type of FLMs has drawn an increasing attention recently in the functional and longitudinal data analysis. Hoover et al. (1998) study the estimation of the regression coefficient functions using the smoothing spline and the local polynomial regression. Fan and Zhang (2000) propose a two-step kernel smoothing procedure for the data collected at the same scheduled time points for each subject. Eubank et al. (2004) develop Bayesian prediction intervals via the smoothing spline for the coefficient curves. Yao et al. (2005) study a special case of the model (1.1), the mean functional model, for the irregular and sparse longitudinal data. To study the possible nonlinear relationship rather than model (1.1), Ferraty et al. (2012) take a fully nonparametric approach.

The FLMs are useful for the functional data when all subjects obey the same linear relationship between the response and covariates. However, in some applications, the subjects might come from an inhomogeneous population which consists of several homogeneous subpopulations/clusters, but within each subpopulation/cluster the concurrent linear model still holds. For this type of applications, a single FLM (1.1) is no longer adequate for modeling the functional data. This motivates us to extend the FLM (1.1) to heterogeneous functional data and propose a new class of functional regression models, called mixtures of FLMs. We will explain in more detail the new model in the next section.

The motivation for our new model also comes from the analysis of a CO_2-GDP dataset. The CO_2-GDP dataset contains two related variables for 175 countries for the years 1980–2005, the CO_2 emissions per capita and the GDP per capita. It is of our interest to model how the CO_2 emissions level depends on the GDP. Fig. 1(a) depicts the CO_2 emission curves of the 175 countries. Fig. 1(b) shows a scatter plot of the cross-sectional subset of the data for the year 2005, along with two fitted regression lines, which demonstrates two different economic development paths among 175 countries. The slope for each subgroup gives the average associated increment of CO_2 emissions per capita given an unit increment in GDP per capita (Huang and Yao, 2012). Most developed countries are from the lower component which has a smaller slope, and the representatives include the United States, the United Kingdom, Canada, Australia, etc. Representative countries from the top component, which has a bigger slope, include Kuwait, Saudi Arabia, Qatar, etc. For the original functional data, it is tempting to fit a two-component mixture of linear regressions to the cross-sectional subset of each year, and check whether the slopes of two components vary over time. However, this naive fitting-one-mixture-a-year procedure suffers serious drawbacks. First, the naive procedure cannot ensure a consistent labeling for each country across different Second, the slope estimates might change rapidly from year to year, since all slopes are estimated separately for different years. Third, it neglects any possible within-subject correlations across years for each country. Thus, a novel estimation procedure needs to be developed for the analysis of the functional CO_2-GDP data to incorporate the information across different years, which is the main purpose of this paper. Our analysis of CO_2-GDP data is an application of functional data modeling in the study of climate change. Functional data analysis has found a broad range of applications, such as economics, biomedicine, geology, environmetrics, paleoclimatology, etc. For example, a nonparametric functional regression model is applied to analyze and forecast the maximum ozone concentration (Aneiros-Pérez et al., 2004), bivariate splines are employed to study the ozone concentration forecasting (Ettinger et al., 2012). Both applications have a real response and functional covariates.

In this article, we propose a mixture of FLMs for heterogeneous functional data. The regression coefficients and covariances for each component are assumed to be smooth functions of t. We first establish the identifiability result for the proposed mixture model under mild conditions. To the best of our knowledge, this is the first identifiability result for mixtures of FLMs so far. We next develop an estimation procedure for the regression coefficients and covariance functions by combining the techniques of the kernel regression, the functional principal component analysis, and the EM algorithm. To choose the number of components for the mixture of FLMs based on the traditionally used information criteria, such as BIC and AIC, we propose an effective degree of freedom for both nonparametric regression coefficients and nonparametric covariance functions by adapting the idea of Fan et al. (2001). The Wilk's type of phenomenon (Fan et al., 2001) is also investigated for the model inference, and a conditional bootstrap method is proposed for the standard error estimation and the hypothesis testing. Finally, we examine the performance of the proposed estimation procedure and the bootstrap method empirically via a Monte Carlo simulation study and an application of the CO_2-GDP data.

The rest of the paper is organized as follows. In Section 2, we introduce the mixture of functional linear models and prove its identifiability. In addition, we develop an estimation procedure by combining the techniques of kernel regression, functional principal component analysis, and EM algorithm. Model selection and the inference are studied in Section 3. Simulation results and a real data application are presented in Section 4. In Section 5, we provide some concluding remarks and discussions. Technical proofs are given in the Appendix.

2. Mixtures of functional linear models

2.1. Model specification and identifiability

In this section we begin with the formulation of the mixture of FLMs. Let $\mathcal{C}$ be a latent class variable with a discrete distribution $P(\mathcal{C} = c) = \pi_c$, for $c = 1, 2, \ldots, C$. We assume that the number of components C is known for the time being, and will discuss how to select C adaptively in the next section. Conditioning on $\mathcal{C} = c$, $\{y(t), t \in T\}$ follows a functional linear model

$$y(t)|_{\mathcal{C}=c} = \mathbf{X}(t)^T \boldsymbol{\beta}_c(t) + \varepsilon_c(t), \tag{2.1}$$

where $\mathbf{X}(t)$ is a p-dimensional random covariate process, $\boldsymbol{\beta}_c(t)$ is an unknown smooth regression coefficient function for the cth component, and $\varepsilon_c(t)$ is a zero mean Gaussian process that is independent of $\mathbf{X}(t)$. We further assume that the error process $\varepsilon_c(t)$ consists of two parts, a trajectory effect $\zeta_c(t)$ and a measurement error effect $e(t)$:

$$\varepsilon_c(t) = \zeta_c(t) + e(t).$$

The trajectory process $\zeta_c(t)$ is independent of $e(t)$ and has a covariance function $\Gamma_c(s, t) = \mathrm{Cov}\{\zeta_c(s), \zeta_c(t)\}$, which is a positive definite smooth function of s and t; and the measurement error $e(t)$ is an uncorrelated process with a constant variance function $\sigma^2(t) \equiv \sigma^2$. The unconditional model of $\{y(t) : t \in T\}$, without knowing the latent variable $\mathcal{C}$, is referred to as the mixture of functional linear models.

When there are no real covariates in the data, and the error processes are Gaussian, the mixture model (2.1) reduces to mixtures of Gaussian processes, which are also known as functional clustering models (James and Sugar, 2003; Luan and Li, 2003; Heard et al., 2006; Ma and Zhong, 2008). Huang et al. (2014) imposes smooth structures for both mean and covariance functions in the mixture of Gaussian processes, and develops estimation procedures based on the kernel regression. For the functional data with real covariates, Yao et al. (2011) propose a mixture of non-concurrent FLMs. Lu and Song (2012) study a mixture of varying coefficient models for the longitudinal data analysis.

Identifiability is of fundamental importance for traditional parametric mixture models, since the parameter estimation and the inference must be based on identifiable models. For a classical finite mixture of parametric models, the identifiability issue is studied in Titterington et al. (1985). Next, we will first give a formal definition of the identifiability for mixtures of FLMs and then prove that the mixture of FLMs is identifiable under mild conditions. To best of our knowledge, this is the first identifiability result for mixtures of FLMs so far.

Definition 1. The mixture of functional linear models (2.1) is said to be identifiable if it does not admit another representation, i.e., if there is another latent variable $\mathcal{G}$, $P(\mathcal{G} = g) = \lambda_g$, $g = 1, \ldots, G$, such that given $\mathcal{G} = g$, $\{Y(t), t \in T\}$ follows a Gaussian process with mean function $\mathbf{X}(t)^T \boldsymbol{\gamma}_g(t)$ and covariance function $\mathrm{Cov}\{Y(s), Y(t)\} = \Omega_g(s, t)$, then $G = C$ and

$$\lambda_g = \pi_g, \qquad \boldsymbol{\gamma}_g(t) = \boldsymbol{\beta}_g(t), \qquad \Omega_g(s, t) = \Gamma_g(s, t), \quad s, t \in T,\ g = 1, \ldots, C,$$

up to the permutation of component labels.

Theorem 1. *Suppose that for each $t \in T$, the domain of $\mathbf{X}(t)$ contains an open set, and that for any $c = 1, \ldots, C$, $\Gamma_c(s, t)$ is a positive definite and bivariate smooth function of s and t, and that $\boldsymbol{\beta}_c(t)$ is a smooth function of t. Let $\mathbf{S} = \{t \in T : (\boldsymbol{\beta}_i(t), \Gamma_i(t, t)) = (\boldsymbol{\beta}_j(t), \Gamma_j(t, t))$ for some $i \neq j$, $1 \leq i, j \leq C\}$. If the set $T \setminus \mathbf{S}$ is not empty, then the above proposed mixture of FLMs is identifiable.*

Based on the above Theorem, the mixture of FLMs is identifiable under some mild conditions. The proof is relegated to the Appendix.

2.2. Estimation procedure

By the well-known Karhunen–Loève theorem (Sapatnekar, 2011), conditioning on $\mathcal{C} = c$, an observed curve $\{y_i(t), \mathbf{X}_i(t)\}$ can be represented as

$$y_i(t) = \mathbf{X}_i(t)^T \boldsymbol{\beta}_c(t) + \sum_{q=1}^{\infty} \xi_{iqc} v_{qc}(t) + e_i(t), \tag{2.2}$$

where $v_{qc}(\cdot)$s are eigenfunctions of the covariance function $\Gamma_c(s,t)$ with corresponding eigenvalues λ_{qc}s, and ξ_{iqc}s are uncorrelated functional principal component (FPC) scores of the trajectories $\zeta_c(t)$ satisfying $\mathrm{E}(\xi_{iqc}) = 0, \mathrm{Var}(\xi_{iqc}) = \lambda_{qc}, \lambda_{1c} \geq \lambda_{2c} \geq \cdots$, and $\sum_{q=1}^{\infty} \lambda_{qc} < \infty$.

Suppose subject $y_i(t)$ is observed at time $t_{ij}, j = 1, \ldots, N_i$. We define the notation $y_{ij} = y_i(t_{ij})$ for the simplicity, and similarly define notations $\varepsilon_{cij}, e_{ij}$, etc. Based on (4.1), conditioning on $\mathcal{C} = c$, the observations $y_{ij}, j = 1, \ldots, N_i$ and $i = 1, \ldots, n$, can be written as

$$y_{ij} = \mathbf{X}_i(t_{ij})^T \boldsymbol{\beta}_c(t_{ij}) + \sum_{q=1}^{\infty} \xi_{iqc} v_{qc}(t_{ij}) + e_{ij}, \tag{2.3}$$

where e_{ij}s are independent and identically distributed as $N(0, \sigma^2)$.

Next, we introduce two procedures for estimating the π_c, $\boldsymbol{\beta}_c(\cdot)$, and the covariance structure. In the first estimation procedure, we pretend that the observations are uncorrelated by ignoring the correlation structure among the data. The idea of using working independence correlation structure has been traditionally used by generalized estimating equation in the longitudinal data analysis (Liang and Zeger, 1986; Lin and Carroll, 2000). Let $\sigma_c^{*2}(t) = \Gamma_c(t,t) + \sigma^2$. It follows that

$$y_{ij} = \mathbf{X}_i(t_{ij})^T \boldsymbol{\beta}_c(t_{ij}) + \varepsilon_{ij}^*, \tag{2.4}$$

where ε_{ij}^* are independent errors satisfying $E(\varepsilon_{ij}^*) = 0$ and $\mathrm{Var}(\varepsilon_{ij}^*) = \sigma_c^{*2}(t_{ij})$. Hence, y_{ij} can be considered coming from the following mixture of Gaussian process:

$$y(t) \sim \sum_{c=1}^{C} \pi_c N\{\mathbf{X}(t)^T \boldsymbol{\beta}_c(t), \sigma_c^{*2}(t)\}. \tag{2.5}$$

Let $\phi(y|\mu, \sigma^2)$ be the density function of $N(\mu, \sigma^2)$. The log-likelihood function of (2.5) is

$$\sum_{i=1}^{n} \log \left[\sum_{c=1}^{C} \pi_c \prod_{j=1}^{N_i} \phi \left\{ y_{ij} | \mathbf{X}_i(t_{ij})^T \boldsymbol{\beta}_c(t_{ij}), \sigma_c^{*2}(t_{ij}) \right\} \right]. \tag{2.6}$$

We next introduce an EM-type estimation procedure to estimate the π_c, $\boldsymbol{\beta}_c(\cdot)$, and $\sigma_c^{*2}(\cdot)$. The derivation of the procedure is given in the Appendix.

Estimation procedure 1 (Working Independence Covariance Structure)

1. *Initial Value*: For the pooled data, fit a C-component mixture of linear regression models with constant proportions and variances and obtain the estimates $\bar{\boldsymbol{\beta}}_c$, $\bar{\sigma}_c^2$ and $\bar{\pi}_c$. Set the initial values $\boldsymbol{\beta}_c^{(1)}(t) = \bar{\boldsymbol{\beta}}_c$, $\sigma_c^{2(1)}(t) = \bar{\sigma}_c^2$, and $\pi_c^{(1)} = \bar{\pi}_c$, $c = 1, \ldots, C$.
2. E-step: For $i = 1, \ldots, n$, and $c = 1, \ldots, C$, calculate

$$r_{ic}^{(l+1)} = \frac{\pi_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij} | X_i(t_{ij})^T \boldsymbol{\beta}_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\} \right]}{\sum_{c=1}^{C} \pi_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij} | X_i(t_{ij})^T \boldsymbol{\beta}_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\} \right]}. \tag{2.7}$$

3. M-step: Update the component proportions by

$$\pi_c^{(l+1)} = \frac{1}{n} \sum_{i=1}^{n} r_{ic}^{(l+1)}. \tag{2.8}$$

For any time t_0 from a set of grid points $\{u_1, \ldots, u_{n_{grid}}\}$, update the slope and variance functions by

$$\boldsymbol{\beta}_c^{(l+1)}(t_0) = \left\{\sum_{i=1}^{n} \mathbf{X}_i^T W_{ic}^{(l+1)}(t_0)\mathbf{X}_i\right\}^{-1} \left\{\sum_{i=1}^{n} \mathbf{X}_i^T W_{ic}^{(l+1)}(t_0)\mathbf{y}_i\right\}, \tag{2.9}$$

$$\sigma_c^{*2(l+1)}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{cij}^{(l+1)}\{y_{ij} - X_i(t_{ij})^T \boldsymbol{\beta}_c^{(l+1)}(t_0)\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{cij}^{(l+1)}}, \tag{2.10}$$

where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iN_i})^T$, $\mathbf{X}_i = (\mathbf{X}_i(t_{i1}), \ldots, \mathbf{X}_i(t_{iN_i}))^T$, $w_{cij}^{(l+1)} = r_{ic}^{(l+1)} K_{h_\beta}(t_{ij} - t_0)$, $W_{ic}^{(l+1)}(t_0) = \mathrm{diag}\{r_{ic}^{(l+1)} K_{h_\beta}(t_{i1} - t_0), \ldots, r_{ic}^{(l+1)} K_{h_\beta}(t_{iN_i} - t_0)\}$, $K_h(\cdot) \equiv h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel density, and h_β is a bandwidth parameter.

4. Iteratively update the E-step and the M-step with $l = 2, 3, \ldots,$ until the algorithm converges. Denote the resulting estimates of π_c, $\boldsymbol{\beta}_c(\cdot)$, and $\sigma_c^{*2}(\cdot)$ by $\hat{\pi}_c$, $\hat{\boldsymbol{\beta}}_c(\cdot)$, and $\hat{\sigma}_c^{*2}(\cdot)$, respectively. Denote the resulting posterior probability as $\hat{r}_{ic}$.

Note that the above estimate does not use the correlation information among the data and thus might lose some efficiency. Next, we introduce an improved estimation procedure which can incorporate the covariance functions into the model estimation by combining the techniques of the kernel regression, the functional principal component analysis (PCA), and the EM algorithm.

Given the estimates from Estimation procedure 1, the covariance function $\Gamma_c(s, t)$ could be estimated by a two-dimensional kernel smoother, which is to minimize

$$\sum_{i=1}^{n} \hat{r}_{ic} \sum_{1\le j\ne l\le N_i} [\hat{\gamma}_{ic}(t_{ij}, t_{il}) - \beta_0]^2 K_{h_\Gamma}(t_{ij} - s) K_{h_\Gamma}(t_{il} - t), \tag{2.11}$$

with respect to β_0, where $\hat{\gamma}_{ic}(t_{ij}, t_{il}) = \{y_{ij} - \mathbf{X}_i(t_{ij})^T \hat{\boldsymbol{\beta}}_c(t_{ij})\}\{y_{il} - \mathbf{X}_i(t_{il})^T \hat{\boldsymbol{\beta}}_c(t_{il})\}$, and h_Γ is a bandwidth parameter for the covariance smoothing. If we can estimate the ξ_{iqc} and $v_{qc}(\cdot)$ in (2.3) from the estimated covariance function $\hat{\Gamma}_c(s, t)$, then we can transfer the correlated data into uncorrelated ones based on (2.3). The estimates of eigenvalues $\hat{\lambda}_{qc}$ and eigenfunctions $\hat{v}_{qc}(\cdot)$ are determined by the equations

$$\int_T \hat{\Gamma}_c(s, t)\hat{v}_{qc}(s)ds = \hat{\lambda}_{qc}\hat{v}_{qc}(t), \tag{2.12}$$

where $\hat{v}_{qc}(t)$ satisfies $\int_T \hat{v}_{qc}^2(t)dt = 1$, and $\int_T \hat{v}_{pc}(t)\hat{v}_{qc}(t)dt = 0$ if $p \ne q$. The above estimation can be implemented by discretizing the covariance estimate $\hat{\Gamma}_c(s, t)$ (Rice and Silverman, 1991). The functional principal component score ξ_{iqc} can then be estimated by

$$\hat{\xi}_{iqc} = \int_T \{y_i(t) - \mathbf{X}_i(t)^T \hat{\boldsymbol{\beta}}_c(t)\}\hat{v}_{qc}(t)dt. \tag{2.13}$$

Let

$$\hat{y}_c(t_{ij}) = y_{ij} - \sum_q \hat{\xi}_{iqc} I(\hat{\lambda}_{qc} > 0)\hat{v}_{qc}(t_{ij}). \tag{2.14}$$

Then, conditioning on $\mathcal{C} = c$, model (2.3) can be approximated by

$$\hat{y}_c(t_{ij}) \approx \mathbf{X}_i(t_{ij})^T \boldsymbol{\beta}_c(t_{ij}) + e_{ij}, \tag{2.15}$$

where e_{ij}'s are independent and identically distributed as $N(0, \sigma^2)$. Hence, using functional PCA, we can transform the correlated data to the uncorrelated one. Based on $\{\hat{y}_c(t_{ij}), i = 1, \ldots, n, j = 1, \ldots, N_i, c = 1, \ldots, C\}$ and (2.15), Estimation procedure 1 can be applied to further improve the estimates of π_c, $\boldsymbol{\beta}_c(\cdot)$, and $\sigma_c(\cdot)$.

Based on the above discussion, we propose the following improved estimation procedure to incorporate the covariance functions to the model estimation.

Estimation procedure 2 (General Covariance Structure)

1. Calculate $\hat{\boldsymbol{\beta}}_c(\cdot)$, $\hat{\pi}_c$, and $\hat{r}_{ic}$ using Estimation procedure 1, and obtain $\hat{y}_c(t_{ij})$ via (2.11)–(2.14). Let

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{n} N_i} \sum_{i=1}^{n}\sum_{c=1}^{C}\sum_{j=1}^{N_i} \hat{r}_{ic}\{\hat{y}_c(t_{ij}) - \mathbf{X}_i(t_{ij})^T \hat{\boldsymbol{\beta}}_c(t_{ij})\}^2.$$

Then set the initial values $\boldsymbol{\beta}_c^{(1)}(\cdot) = \hat{\boldsymbol{\beta}}_c(\cdot)$, $\pi_c^{(1)} = \hat{\pi}_c$, $r_{ic}^{(1)} = \hat{r}_{ic}$, $c = 1, \ldots, C$, and $\sigma^{2(1)} = \hat{\sigma}^2$.

2. Estimate the covariance function $\Gamma_c(s,t)$ by

$$\Gamma_c^{(l+1)}(s,t)=\frac{\sum_{i=1}^{n} r_{ic}^{(l)} \sum_{1\le j\ne l\le N_i} \gamma_{ic}^{(l)}(t_{ij},t_{il}) K_{h_\Gamma}(t_{ij}-s) K_{h_\Gamma}(t_{il}-t)}{\sum_{i=1}^{n} r_{ic}^{(l)} \sum_{1\le j\ne l\le N_i} K_{h_\Gamma}(t_{ij}-s) K_{h_\Gamma}(t_{il}-t)}, \tag{2.16}$$

where $\gamma_{ic}^{(l)}(t_{ij},t_{il})=\{y_{ij}-\mathbf{X}_i(t_{ij})^T\boldsymbol{\beta}_c^{(l)}(t_{ij})\}\{y_{il}-\mathbf{X}_i(t_{il})^T\boldsymbol{\beta}_c^{(l)}(t_{il})\}$. Let $\lambda_{qc}^{(l+1)}$ and $v_{qc}^{(l+1)}(\cdot)$ be the estimated eigenvalues and eigenfunctions, respectively, from the covariance estimate $\Gamma_c^{(l+1)}(s,t)$. To ensure positive-definiteness of the covariance estimates, we set $\Gamma_c^{(l+1)}(s,t)=\sum_q \lambda_{qc}^{(l+1)} I(\lambda_{qc}^{(l+1)}>0) v_{qc}^{(l+1)}(s) v_{qc}^{(l+1)}(t)$.

3. Calculate the transformed response

$$y_c^{(l+1)}(t_{ij})=y_{ij}-\sum_q \xi_{iqc}^{(l+1)} I(\lambda_{qc}^{(l+1)}>0) v_{qc}^{(l+1)}(t_{ij}),$$

where

$$\xi_{iqc}^{(l+1)}=\int_T \{y_i(t)-\mathbf{X}_i(t)^T\boldsymbol{\beta}_c^{(l)}(t)\} v_{qc}^{(l+1)}(t)dt. \tag{2.17}$$

4. One cycle E-step:

$$r_{ic}^{(l+1)}=\frac{\pi_c^{(l)}\left[\prod_{j=1}^{N_i}\phi\{y_c^{(l+1)}(t_{ij})|\mathbf{X}_i(t_{ij})^T\boldsymbol{\beta}_c^{(l)}(t_{ij}),\sigma^{2(l)}\}\right]}{\sum_{c=1}^{C}\pi_c^{(l)}\left[\prod_{j=1}^{N_i}\phi\{y_c^{(l+1)}(t_{ij})|\mathbf{X}_i(t_{ij})^T\boldsymbol{\beta}_c^{(l)}(t_{ij}),\sigma^{2(l)}\}\right]}. \tag{2.18}$$

5. One cycle M-step: For $t_0\in\{u_1,\cdots,u_{n_{grid}}\}$,

$$\pi_c^{(l+1)}=\frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l+1)}, \tag{2.19}$$

$$\boldsymbol{\beta}_c^{(l+1)}(t_0)=\left\{\sum_{i=1}^{n}\mathbf{X}_i^T W_{ic}^{(l+1)}(t_0)\mathbf{X}_i\right\}^{-1}\left\{\sum_{i=1}^{n}\mathbf{X}_i^T W_{ic}^{(l+1)}(t_0)\mathbf{y}_{ic}^{(l+1)}\right\}, \tag{2.20}$$

$$\sigma^{2(l+1)}=\frac{1}{\sum_{i=1}^{n}N_i}\sum_{i=1}^{n}\sum_{c=1}^{C}\sum_{j=1}^{N_i} r_{ic}^{(l+1)}\{y_c^{(l+1)}(t_{ij})-\mathbf{X}_i(t_{ij})^T\boldsymbol{\beta}_c^{(l+1)}(t_{ij})\}^2, \tag{2.21}$$

where

$$\mathbf{y}_{ic}^{(l+1)}=\{y_c^{(l+1)}(t_{i1}),\ldots,y_c^{(l+1)}(t_{iN_i})\}^T,$$
$$W_{ic}^{(l+1)}(t_0)=\mathrm{diag}\{r_{ic}^{(l+1)}K_{h_\beta}(t_{i1}-t_0),\ldots,r_{ic}^{(l+1)}K_{h_\beta}(t_{iN_i}-t_0)\}.$$

6. Iterate Steps 2–5 with $l=2,3,\ldots,$ until convergence.

3. Model selection and inference

3.1. Model selection

The model selection includes selection of the model type (the full model (2.3) and the reduced model (2.4)), the number of components C, the bandwidths h_β and h_Γ, and the number of eigenfunctions. We discuss these aspects in this section.

Selection of the number of components C is a difficult yet important issue for mixture models. Many efforts have been made to solve this problem for parametric mixture models (Hartigan, 1985; Chen et al., 2001; Li and Chen, 2010; Leroux, 1992; Frühwirth-Schnatter, 2006; McLachlan and Peel, 2000). Bayesian information criterion (BIC) is one of the most extensively used methods. The BIC has the form $-2\mathcal{L}+\log(n)\times df$, where $\mathcal{L}$ is the maximum log-likelihood function, and df is the model degree of freedom which measures the complexity of the model. However, it is not clear how to define the model complexity for the mixture of FLMs, since it contains one-dimensional mean functions smoothing, and the two-dimensional covariance functions smoothing. Next, we propose to adopt the idea of Fan et al. (2001) to define the effective degree of freedom for nonparametric kernel smoothing and then apply the traditional information criteria to do model selection.

For the one-dimensional mean functions smoothing, based on Fan et al. (2001), we define the effective degree of freedom by

$$df_\beta = \tau_K h_\beta^{-1}|\Omega| \left\{ K(0) - \frac{1}{2}\int K^2(t)dt \right\},$$

where Ω is the support of the time t, and

$$\tau_K = \frac{K(0) - \frac{1}{2}\int K^2(t)dt}{\int \{K(t) - \frac{1}{2}K * K(t)\}^2 dt}.$$

Note that the df_β depends on $h_\beta^{-1}|\Omega|$. As discussed in Remark 3.2 in Fan et al. (2001), the number of parameters is $h_\beta^{-1}|\Omega|$ if the one dimensional support is partitioned into intervals of length h_β, and piecewise constant functions are used for approximation. Similarly, for the two-dimensional covariance functions smoothing, we can define the effective degree of freedom as

$$df_\Gamma = \tau_K^2 h_\Gamma^{-2}|\Omega|^2 \left\{ K(0) - \frac{1}{2}\int K^2(t)dt \right\}^2.$$

Based on the above definitions, the degree of freedom for the model (2.4) is $(pC + C) \times df_\beta + C - 1$, and the degree of freedom for the model (2.3) is $p \times C \times df_\beta + C \times df_\Gamma + C$. We choose the model with the minimum BIC in the candidate set which consists of the model (2.3) and the model (2.4) with different C. Note that the degree of freedom depends on both C and bandwidths. In practice, we can apply BIC for a wide range of bandwidths.

Once the number of components C is determined, we need to choose the bandwidths and the number of eigenfunctions. Choosing the bandwidths has long been a difficult problem for nonparametric and semiparametric models. For a comprehensive review on the bandwidth selection in the kernel regression, see, for example, Marron (1988) and Fan and Gijbels (1996). In this paper, we consider a multifold cross-validation (CV) method to choose the bandwidths. In the model (2.4), a conventional CV is applied as we use the same bandwidth for both the regression functions and covariance functions for the simplicity. For the model (2.3), we need to select both h_β and h_Γ. Hence, CV needs to be performed in a two-dimensional domain. The simulation results in Section 4 demonstrate that the proposed estimation procedure works well for a wide range of bandwidths.

Given selected bandwidths in the model (2.3), the number of eigenfunctions might be chosen using one-curve-leave-out CV and pseudo-AIC criterion (Rice and Silverman, 1991; Yao et al., 2005). From our simulation experience, these methods do not work very well in our model setting. Following Huang et al. (2014), we choose the number of eigenfunctions by an empirical criterion such that the percentage of total variation explained is above certain percentage, such as 95%.

3.2. Model inference

For the proposed model, it is of interest to test whether the coefficient functions $\boldsymbol{\beta}_c(t)$s actually depend on t. This leads to the following hypothesis testing problem:

$$H_0 : \boldsymbol{\beta}_c(t) \equiv \boldsymbol{\beta}_c, \quad c = 1, \ldots, C. \tag{3.1}$$

Let $\ell(H_0)$ and $\ell(H_1)$ be the maximum log-likelihoods under null and alternative hypotheses, respectively. Then we construct a likelihood ratio test statistic

$$T = 2\{\ell(H_1) - \ell(H_0)\}.$$

Since both the null and alternative models are semiparametric ones, such hypothesis test belongs to the Wilks phenomenon and generalized likelihood ratio theory (Fan et al., 2001) for the semiparametric modeling. In this paper, we shall first demonstrate that the Wilk's type of results hold for our model via a Monte Carlo simulation study, i.e., the null distribution of (3.1) does not depend on the nuisance parameters of the null model. Then we apply a conditional bootstrap method (Cai et al., 2000; Fan et al., 1999) to estimate the null distribution. See Section 4 for more detail about its implementation. The conditional bootstrap method will also be used for the standard error estimation and to construct pointwise confidence intervals (CIs) for coefficient functions.

4. Simulation and application

To measure the performance of our estimators, we use the root of the average squared errors (RASE). For the estimate of the parameter π, define $\text{RASE}_\pi^2 = \sum_{c=1}^{C-1}\{\hat{\pi}_c - \pi_c\}^2$. For the estimates of the regression coefficient functions $\boldsymbol{\beta}_c(t)$, define

$$\text{RASE}_\beta^2 = n_{grid}^{-1} \sum_{c=1}^{C} \sum_{j=1}^{n_{grid}} \|\hat{\boldsymbol{\beta}}_c(u_j) - \boldsymbol{\beta}_c(u_j)\|^2,$$

where $\{u_j, j = 1, \ldots, n_{grid}\}$ are a set of grid points. The number of grid points is set to be $n_{grid} = 50$, and all the grid points are evenly distributed on the range of t.

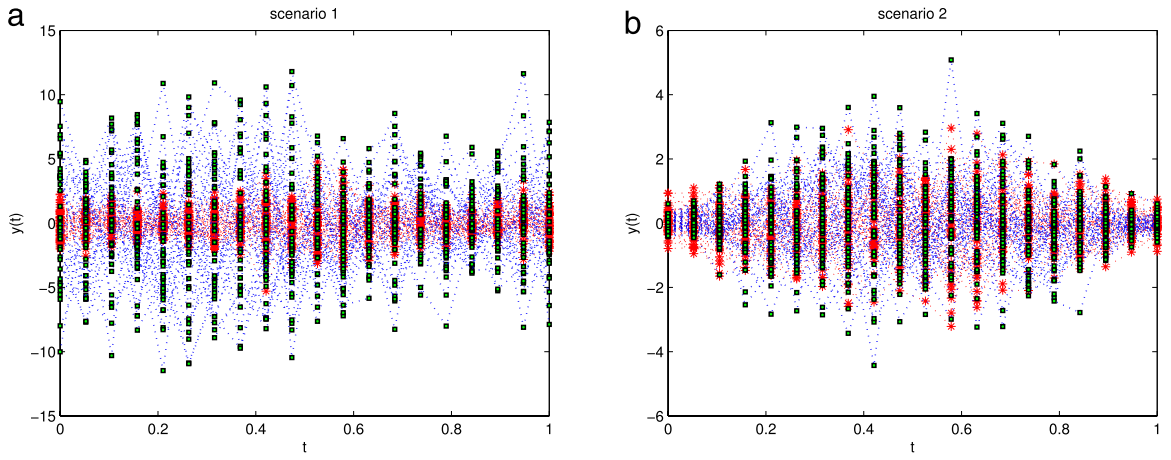

Fig. 2. (a) Response of a typical sample for the well-separated setting; (b) Response of typical sample for heavy-overlap setting.

4.1. Simulation study

We shall consider two simulation scenarios as follows.

Scenario 1

$$\pi_1 = 0.6, \quad \pi_2 = 1 - \pi_1 = 0.4, \quad \text{and} \quad \sigma^2 = 0.25,$$
$$\boldsymbol{\beta}_1(t) = (\sin(\pi t), \cos(2\pi t)) \quad \text{and} \quad \boldsymbol{\beta}_2(t) = (t^2 - 3, \sin(2\pi t) + 3),$$
$$v_{11}(t) = \sqrt{2}\sin(4\pi t) \quad \text{and} \quad v_{21}(t) = \sqrt{2}\cos(4\pi t),$$
$$v_{12}(t) = \sqrt{2}\sin(\pi t) \quad \text{and} \quad v_{22}(t) = \sqrt{2}\cos(\pi t),$$
$$\lambda_{1c} = 0.04, \quad \lambda_{2c} = 0.01 \quad \text{and} \quad \lambda_{qc} = 0, \quad \text{for } q > 2,\ c = 1, 2.$$

Scenario 2

$$\pi_1 = 0.45, \quad \pi_2 = 1 - \pi_1 = 0.55, \quad \text{and} \quad \sigma^2 = 0.25,$$
$$\boldsymbol{\beta}_1(t) = (0, \sin(\pi t)) \quad \text{and} \quad \boldsymbol{\beta}_2(t) = (0, 1.5\sin(\pi t)),$$
$$v_{11}(t) = \sqrt{2}\sin(4\pi t) \quad \text{and} \quad v_{21}(t) = \sqrt{2}\cos(4\pi t),$$
$$v_{12}(t) = \sqrt{2}\sin(\pi t) \quad \text{and} \quad v_{22}(t) = \sqrt{2}\cos(\pi t),$$
$$\lambda_{11} = 0.16, \quad \lambda_{21} = 0.04 \quad \text{and} \quad \lambda_{q1} = 0, \quad \text{for } q > 2,$$
$$\lambda_{12} = 0.04, \quad \lambda_{22} = 0.01 \quad \text{and} \quad \lambda_{q2} = 0, \quad \text{for } q > 2.$$

In the first scenario, the two components are well-separated; while in the second scenario, the two components heavily overlap. In both scenarios, the two components have different correlation structures. For each scenario, the simulated data of size $n = 100$ are observed at grid points $\{k/N, k = 1, \ldots, N\}$ for both components, where N is set to be 20 and 40, respectively. At each grid point, the predictor X is generated from a one-dimensional standard normal distribution. The principal component scores ξ_{iqc} are generated from $N(0, \lambda_{qc})$, $q = 1, 2$, and $c = 1, 2$. Typical samples of the response from the two scenarios are depicted in Fig. 2.

Notations M_F and M_R stand for the full model (2.3) with the estimation procedure 2, and the reduced model (2.4) with the estimation procedure 1, respectively. The Epanechnikov kernel is used for the kernel smoothing in estimation.

We first test the performance of the proposed model selection method based on BIC and the defined effective degree of freedom for BIC. To illustrate the method, we design a contrast scenario 1b which is the same as scenario 1 except that it has an independent covariance structure.

Scenario 1b

$$\pi_1 = 0.6, \quad \pi_2 = 1 - \pi_1 = 0.4,$$
$$\boldsymbol{\beta}_1(t) = (\sin(\pi t), \cos(2\pi t)) \quad \text{and} \quad \boldsymbol{\beta}_2(t) = (t^2 - 3, \sin(2\pi t) + 3),$$
$$\sigma_1^2(t) = 0.2\sin(\pi t) + 0.25 \quad \text{and} \quad \sigma_2^2(t) = 0.3\sin(\pi t) + 0.25.$$

For both scenarios 1 and 1b, we fit the mixture of FLMs under both the full model (M_F) and the reduced model (M_R) with 1, 2, and 3 components under 9 different pairs of bandwidths where $h_\beta \in \{0.06, 0.08, 0.10\}$ and $h_\Gamma \in \{0.28, 0.35, 0.42\}$, and then compare the corresponding BIC scores. The C is selected by minimizing the BIC scores over M_F and M_R, the three

Table 1
Frequencies of the selected C and the model type by BIC ($N = 20$).

		Scenario 1	Scenario 1b
M_R	C = 1	0	0
	C = 2	6	100
	C = 3	0	0
M_F	C = 1	0	0
	C = 2	94	0
	C = 3	0	0

Table 2
Optimal bandwidth pairs $(\hat{h}_\beta, \hat{h}_\Gamma)$.

Scen.	N	M_R h_β	M_F h_Γ
1	20	0.0805	0.3500
	40	0.0730	0.1290
2	20	0.0650	0.1620
	40	0.0560	0.0800

Table 3
Mean and standard deviation of RASEs.

		M_R			M_F		
Scen.	N	RASE_β	RASE_π	$\pi_1 = 0.6$	RASE_β	RASE_π	$\pi_1 = 0.6$
1	20	0.013(0.003)	0.002(0.003)	0.602(0.050)	0.013(0.003)	0.002(0.003)	0.602(0.050)
	40	0.008(0.002)	0.002(0.003)	0.601(0.047)	0.007(0.001)	0.002(0.003)	0.601(0.047)
Scen.	N	RASE_β	RASE_π	$\pi_1 = 0.45$	RASE_β	RASE_π	$\pi_1 = 0.45$
2	20	0.059(0.288)	0.024(0.071)	0.489(0.149)	0.009(0.043)	0.003(0.014)	0.454(0.057)
	40	0.017(0.063)	0.021(0.068)	0.496(0.139)	0.002(0.014)	0.003(0.004)	0.451(0.051)

values of C, and the 9 pairs of bandwidths (h_β, h_Γ). The frequencies of selected Cs in 100 simulations are presented in Table 1. From these results, we can see that for the data with independent covariance structure (scenario 1b), the proportions for BIC to choose the correct C and the reduced model (2.4) are 100%; while for the data with non-isotropic covariance structure (scenario 1), the proportion of selecting the correct C is 100%, and the proportion of selecting the correct C and the full model (2.3) is 94%. The above results demonstrate the effectiveness of the proposed model selection method. In addition, the new method can not only choose the correct number of components, but also identify the proper model (M_F or M_R) to use.

In the following simulation, we assume that the number of components C is known. For each simulated dataset, we obtain the optimal bandwidths for coefficient and covariance functions using a 5-fold CV method. However, such a CV method performs well at a computational expense. To ease the computation burden, we fix the bandwidth pairs for each simulated dataset. The bandwidth pairs in Table 2 are selected as the average of optimal CV bandwidths of several simulated datasets. From the table, we can see that both h_β and h_Γ decrease as N increases. Table 3 reports the mean and standard deviation of $\text{RASE}_{\boldsymbol{\beta}}$, RASE_π, and the estimated π for both scenarios 1 and 2 over 500 simulations. The results show that M_R and M_F have similar performance and work well with the selected bandwidths in scenario 1. In contrast, in scenario 2, M_F does a much better job than M_R, producing less bias for the parameter π_1, and smaller $\text{RASE}_{\boldsymbol{\beta}}$ and RASE_π. Therefore, when the mixture components are close, incorporating the covariance functions into the model estimation using M_F could significantly improve the accuracy of the model estimation.

Next, we conduct a simulation study to investigate the Wilk's type of phenomenon for the hypothesis testing of coefficient functions. The simulated data are generated according to scenario 2, with the only difference that the coefficient functions are constants. We take three different values of the constant $\boldsymbol{\beta}_c$, $\{(0, -0.5), (0, 0.5)\}$, $\{(0, -0.5), (0, 1)\}$, and $\{(0, 0.5), (0, 1.5)\}$. Since the covariances are non-isotropic, we incorporate the covariance functions in the estimation for both models under the null and alternative hypotheses, and the corresponding log-likelihoods $\ell(H_0)$ and $\ell(H_1)$ are calculated. Let T denote the likelihood ratio test statistic. We compute the unconditional null distributions of T based on 500 Monte Carlo simulations. The resulting three density estimates are very close to each other, plotted as dotted lines in Fig. 3(a). As expected, the asymptotic distribution of T under the null hypothesis is not sensitive to the true values of $\boldsymbol{\beta}_c$. Next, to validate the conditional bootstrap method, we choose three typical samples generated using these three different $\boldsymbol{\beta}_c$. For each typical sample, we first obtain the estimates, and then compute the conditional null distributions of T based on 500 parametric bootstrap samples. The resulting three densities are depicted as solid curves in Fig. 3 (b). From Fig. 3 (b), we can see that the conditional bootstrapped distributions are very close and work reasonably well in approximating the true null distribution.

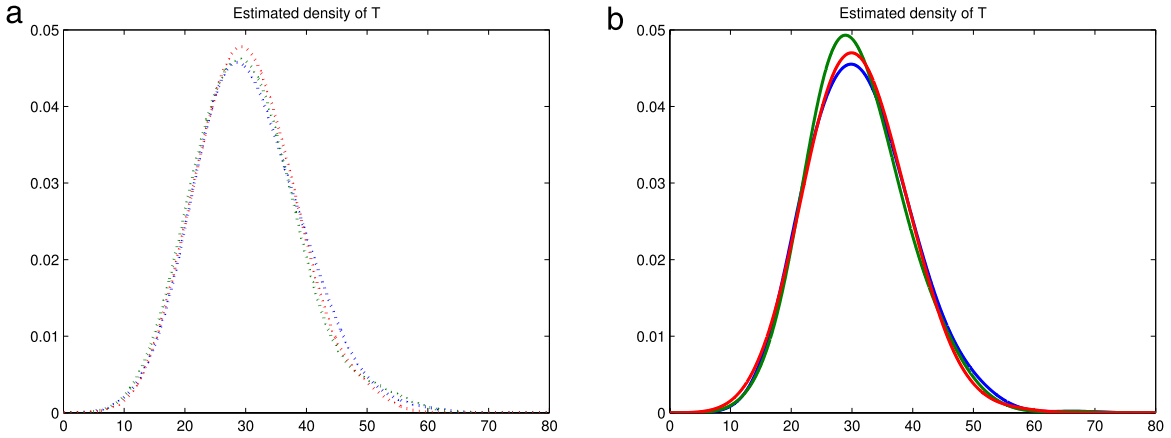

Fig. 3. (a) The estimated densities of unconditional null distributions of T for three different null hypotheses; (b) the estimated densities of conditional null distributions of T (solid lines). The online version of this figure is in color.

Table 4
Standard errors via bootstrap (scenario 1, $N = 20$).

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	SD	0.044	0.044	0.049	0.050	0.047	0.044	0.048	0.047	0.040
$\beta_{11}(\cdot)$	SE	0.042	0.043	0.044	0.045	0.046	0.046	0.044	0.043	0.042
	Std	0.003	0.004	0.004	0.004	0.005	0.005	0.004	0.004	0.004
	SD	0.043	0.051	0.054	0.054	0.051	0.047	0.052	0.050	0.042
$\beta_{12}(\cdot)$	SE	0.042	0.044	0.046	0.046	0.046	0.046	0.045	0.043	0.043
	Std	0.004	0.004	0.005	0.005	0.004	0.005	0.004	0.004	0.004
	SD	0.052	0.056	0.044	0.057	0.058	0.063	0.054	0.050	0.056
$\beta_{21}(\cdot)$	SE	0.053	0.052	0.052	0.055	0.059	0.056	0.054	0.052	0.053
	Std	0.006	0.006	0.005	0.006	0.007	0.006	0.005	0.005	0.005
	SD	0.054	0.054	0.054	0.061	0.062	0.062	0.050	0.057	0.057
$\beta_{22}(\cdot)$	SE	0.054	0.051	0.053	0.058	0.061	0.057	0.052	0.052	0.054
	Std	0.006	0.006	0.006	0.006	0.007	0.006	0.006	0.005	0.005

Table 5
Standard errors via bootstrap (scenario 2, $N = 20$).

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	SD	0.021	0.024	0.021	0.023	0.025	0.024	0.025	0.023	0.026
$\beta_{12}(\cdot)$	SE	0.025	0.026	0.026	0.027	0.025	0.027	0.027	0.027	0.025
	Std	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
	SD	0.026	0.025	0.020	0.022	0.019	0.019	0.020	0.023	0.024
$\beta_{22}(\cdot)$	SE	0.023	0.024	0.023	0.022	0.022	0.022	0.023	0.024	0.023
	Std	0.003	0.003	0.003	0.003	0.002	0.003	0.003	0.003	0.003

Next, we investigate the accuracy of the conditional bootstrap method to estimate the standard errors for the coefficient estimates. We use SE and Std to denote the average and the standard deviation of 200 bootstrapped standard errors, respectively, and use SD to denote the standard deviation of the 200 estimates. SD can be considered as the "real" standard errors, and serves as a benchmark for the comparison. These quantities are given in Tables 4 and 5. From the two tables, we can see that SD and SE are very close, which demonstrates the effectiveness of the conditional bootstrap method as a tool to provide the standard error estimates.

Now we compare our model with an existing alternative procedure, the random effects regression mixture (RERM) model (DeSarbo and Cron, 1988; Verbeke and Lesaffre, 1996; Xu and Hedeker, 2001), with a spline representation for the nonparametric structure. Conditioning on the component membership $\mathcal{C} = c$,

$$y_i(t) = \mathbf{X}_i(t)^T \boldsymbol{\beta}_c(t) + e_i(t), \tag{4.1}$$

where $e(t)$ is an uncorrelated process with a constant variance function $\sigma^2(t) \equiv \sigma^2$. We assume that element functions in $\boldsymbol{\beta}_c(t) = \{\beta_{c1}(t), \ldots, \beta_{cp}(t)\}$ are approximated by the following sum of the spline basis $B_k(t)$:

$$\beta_{cj}(t) = \sum_{k=1}^{K} \gamma_{cjk} B_k(t), \tag{4.2}$$

Table 6
Mean and Standard Deviation of RASEs.

		RERM			M_F		
Scen.	N	RASE_β	RASE_π	$\pi_1 = 0.6$	RASE_β	RASE_π	$\pi_1 = 0.6$
1	20	0.005(0.002)	0.003(0.004)	0.597(0.050)	0.013(0.003)	0.002(0.003)	0.600(0.048)
	40	0.003(0.001)	0.003(0.004)	0.599(0.050)	0.007(0.001)	0.003(0.004)	0.606(0.051)
Scen.	N	RASE_β	RASE_π	$\pi_1 = 0.45$	RASE_β	RASE_π	$\pi_1 = 0.45$
2	20	0.011(0.036)	0.012(0.023)	0.385(0.091)	0.009(0.043)	0.003(0.014)	0.454(0.057)
	40	0.003(0.016)	0.004(0.018)	0.438(0.065)	0.002(0.014)	0.003(0.004)	0.451(0.051)

Table 7
Mean value of the criterion errors on the test sample.

Model	Absolute error	Quadratic error
NFRM	13.050	15.574
M_F	3.388	1.187

for $j = 1, \ldots, p$. Let $\boldsymbol{\gamma}_{cj} = (\gamma_{cj1}, \ldots, \gamma_{cjK})^T$. For the random effects regression mixture (RERM) model with a spline representation, we assume that $\boldsymbol{\gamma}_{cj} \sim N(\boldsymbol{\mu}_{cj}, \mathbf{R}_{cj})$. The unknown parameters in the model are $\{\pi_c, \boldsymbol{\mu}_{cj}, \mathbf{R}_{cj}\}$ for $c = 1, \ldots, C; j = 1, \ldots, p$, under constrains $\pi_c > 0$, and $\sum_{c=1}^{C} \pi_c = 1$. The parameters can be estimated using standard MLE procedure (e.g., EM algorithm) of RERM. Let $\{\hat{\pi}_c, \hat{\boldsymbol{\mu}}_{cj}, \hat{\mathbf{R}}_{cj}\}$ be the estimates. Then the mean function estimation can be expressed as

$$\hat{\beta}_{cj}(t) = \sum_{k=1}^{K} \hat{\mu}_{cjk} B_k(t), \tag{4.3}$$

where $\hat{\mu}_{cjk}$ are elements of $\boldsymbol{\mu}_{cj}$. For convenience, we use a cubic-spline representation with evenly spaced knots. The number of knots is chosen such that the degree of freedom of the RERM is about the same as the effective degree of freedom of our proposed model (see Section 3.1 of our paper). We perform the calculation using R package "mixtools" (see Young et al., 2007).

We summarize the comparison results in Table 6. Notations "RERM" and "M_F" stand for the random effects regression mixture model and our model respectively. The two scenarios are the same as those in Section 4.1. From Table 6, we can see that when the two components are well-separated (Scenario 1), both estimation procedures give similar results for the estimation of π, while RERM has better estimation of $\boldsymbol{\beta}(t)$. When the components are heavily overlapped (Scenario 2), the proposed procedure M_F gives better results compared to RERM.

We also compare our method with the fully nonparametric functional regression model (NFRM) of Ferraty et al. (2012). We use the scenario 2 ($N = 20$) as an illustration. Each dataset is split randomly into a training sample of 75 observations and a testing sample of 25 observations, and the process is replicated 100 times. The training sample is used to select and estimate the parameters of the models and the testing sample is used to compare prediction performance for the two models. The criteria for prediction comparison are the absolute error $AE = \Sigma_i |Y_i - \hat{Y}_i|$ and the quadratic error $QE = \Sigma_i (Y_i - \hat{Y}_i)^2$. Table 7 gives the mean values of AE and QE and shows that our model works better in terms of the prediction errors.

4.2. CO_2-GDP data application

In this section, we analyze the CO_2-GDP data using the proposed model and estimation procedure. The data record the CO_2 emission per capita and the GDP per capita for 175 countries from 1980 to 2005. It has a balanced structure, with one observation each year for each country. The trajectories of the CO_2 emission for all the 175 countries are assembled in Fig. 1(a). Huang and Yao (2012) showed that a cross-sectional subset of the data in year 2005 can be modeled by a 2-component mixture of regression models with varying proportion, with each component revealing a linear effect of the GDP per capita on the CO_2 emission per capita. In this study, we are interested in whether this linear effect of the GDP on the CO_2 emission varies over time for the two components.

We first choose the model type and the number of components C via BIC. Using bandwidths $h_\beta \in \{0.07, 0.09, 0.11\}$ and $h_\Gamma \in \{0.11, 0.13, 0.15\}$, we fit a mixture of FLMs by M_F and M_R to the data with one, two, three, and four components, respectively. Then we calculate and compare their corresponding BIC scores. For these bandwidths, the minimum BIC score is achieved at $C = 2$ with M_F. Hence a two-component mixture model is selected. This agrees with the cross-sectional analysis by Huang and Yao (2012). Our result also indicates that we should use the full model (2.3) and the estimation procedure two to incorporate the correlation structure. With M_F and $C = 2$, a 5-fold CV suggests that the optimal bandwidths for estimating coefficient and covariance functions are 0.085 and 0.13, respectively. With these optimal bandwidths, the 95% percent rule-of-thumb criterion selects 2 eigenfunctions for each component. We then apply the test procedure from Section 3.2 to test whether the coefficient functions change over time. The test statistic T is 571.81 using the estimation procedure 2, with

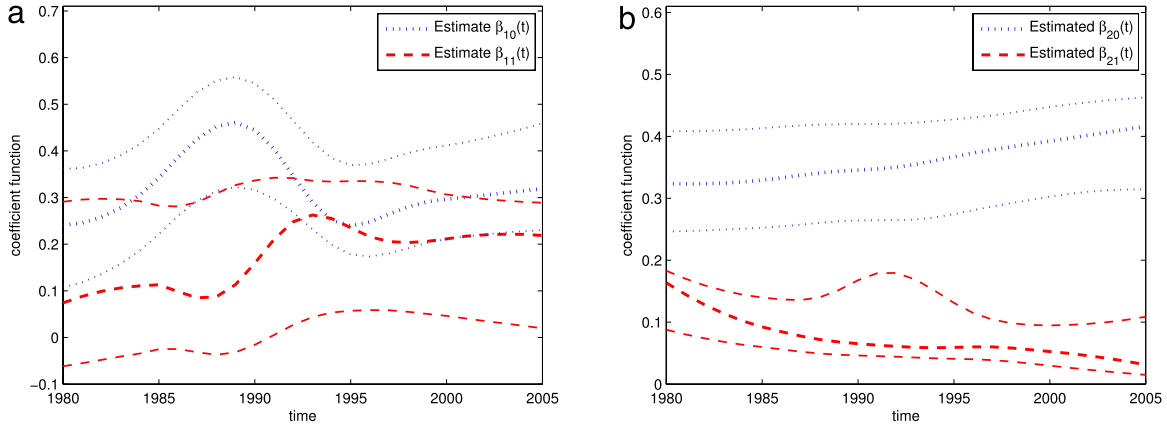

Fig. 4. The estimated coefficient functions with the 95% pointwise confidence intervals for the CO_2-GDP data. (a) Coefficient functions of component 1 ; (b) coefficient functions of component 2.

Table 8
Mean value of the criterion errors on the test sample.

Model	Absolute error	Quadratic error
NFRM	5.375	3.831
M_F	0.712	0.153

p-value (<0.01) being close to zero. Therefore, we reject the null hypotheses of constant coefficient functions, and conclude that a mixture of FLMs with general covariance structure is needed for the analysis of the CO_2-GDP data.

The estimated mixing proportions for the two components are 0.1524 and 0.8476, respectively. We label the component with the proportion 0.1524 as the first component, and the other as the second component. Countries in the first component are those of low GDP per capita and relatively high CO_2 emissions, including Kuwait, United Arab Emirates, Russian Federation, Georgia, etc. Countries in the second component are those of high GDP per capita with relative low CO_2 emissions and thus have healthier economic development path. Representatives in the second component are the United States, Canada, Australia, France, etc. Two estimated coefficient functions $\boldsymbol{\beta}_c(t)$ are shown in Fig. 4, together with their bootstrapped point-wise confidence intervals. Fig. 4(a) demonstrates that the slope function for the first component increases with time for the period of 1980 to 1994, and then remains relatively flat at a high level from 1995 to 2005. In contrast, Fig. 4(b) shows that for countries in the second component, the slope function decreases slowly over the period of 1980–2005. Therefore, for the countries in component 2, the increment of the CO_2 emissions per capita associated with a unit increment of the GDP per capita decreases slowly over the period of 1980–2005, and is lower than that of the component 1 after 1985. Therefore, the countries in the second component had been improving their economic structure along the years. Our result is in agreement with the finding of Garnaut et al. (2008) on emissions/GDP elasticity in general.

We now use NFRM of Ferraty et al. (2012) to analyze the CO_2-GDP data and compare the result with the analysis above. Similar to what we did in simulation, we randomly split the dataset into a training sample of 125 observations and a testing sample of 50 observations, and repeat it 100 times. The training and testing samples are used for parameter estimations and prediction testing respectively. The mean values of *AE* and *QE* are given in Table 8, which demonstrate again the our method performs favorably.

5. Concluding remarks

In this paper, we proposed a new class of mixture of functional linear models to study relationship in inhomogeneous functional data. We showed that the proposed mixture models are identifiable under mild conditions. We developed estimation procedures using the kernel regression, the EM algorithm, and the functional principal component analysis. In order to check whether the coefficient functions are actually varying over time, we employed a semiparametric maximum likelihood ratio test and estimated its null distribution by a conditional bootstrap method. Simulation studies and a real data application demonstrated the effectiveness of the proposed methodology.

The simulated and real data in this paper are observed at regular grid points. For irregular and unbalanced data, one may linearly interpolate the data over a regular grid points, and then apply our estimation procedure. Theoretical properties, such as the consistency and the asymptotic normality, of the proposed estimation procedure have not been established. One might be able to establish these properties in the spirit of the work of Hoshikawa (2013). It requires more research.

Acknowledgments

The authors thank the editor, the associate editor, and reviewers for their constructive comments that have led to a dramatic improvement of the earlier version of this article. Wang's research was partially supported by National Natural Science Foundation of China (NSFC) grant (11371235) and Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (SRF for ROCS, SEM). Huang's research was partially supported by NSFC grant (11301324) and Shanghai Chenguang Program. Yao's research was supported by National Science Foundation (NSF) grant DMS-1461677.

Appendix

Lemma 1. *Consider mixtures of linear regression models with Gaussian errors*

$$Y|\mathbf{x} \sim \sum_{c=1}^{C} \pi_c N(\mathbf{x}^T\boldsymbol{\beta}_c, \sigma_c^2), \tag{A.1}$$

where $\boldsymbol{\beta}_c = (\beta_{0c}, \beta_{1c}, \ldots, \beta_{pc})^T$. *Suppose that* $(\boldsymbol{\beta}_c, \sigma_c^2)$, $c = 1, \ldots, C$, *are distinct pairs, and that the domain* $\mathcal{X}$ *of* $\mathbf{x}$ *contains an open subset in* $\mathbb{R}^p$. *Then the above mixture of linear regression models is identifiable.*

Lemma 1 can be viewed as a special case of Theorem 1 from Huang and Yao (2012) with constant proportions. Hennig (2000) also obtained a similar result earlier.

Lemma 2. *The multivariate Gaussian mixtures are identifiable.*

This lemma is the Proposition 2 from Yakowitz and Spragins (1968).

Proof of Theorem 1. The proof can be done in a similar manner to Huang et al. (2014). Suppose that $\{Y(t), t \in T\}$ admits another representation such that given a latent variable $\mathcal{G} = g$, $\{Y(t), t \in T\}$ follows a Gaussian process with mean function $\mathbf{X}(t)^T\boldsymbol{\gamma}_g(t)$ and covariance function $\mathrm{Cov}\{Y(s), Y(t)\} = \Omega_g(s,t)$, $g = 1, \ldots, G$. In addition, $P(\mathcal{G} = g) = \lambda_g$. Therefore,

$$Y(t) \sim \sum_{g=1}^{G} \lambda_g N(\mathbf{X}(t)^T\boldsymbol{\gamma}_g(t), \qquad \Omega_g(t,t)) = \sum_{c=1}^{C} \pi_c N(\mathbf{X}(t)^T\boldsymbol{\beta}_c(t), \Gamma_c(t,t)).$$

For any fixed $t \in T \setminus \mathbf{S}$, $(\boldsymbol{\beta}_c(t), \Gamma_c(t,t))$, $c = 1, \ldots, C$, are distinct pairs. By Lemma 1, $G = C$, and there exists a permutation $\omega_t = \{\omega_t(1), \ldots, \omega_t(C)\}$ which may depend on t such that

$$\lambda_{\omega_t(c)} = \pi_c, \qquad \boldsymbol{\gamma}_{\omega_t(c)}(t) = \boldsymbol{\beta}_c(t), \qquad \Omega_{\omega_t(c)}(t,t) = \Gamma_c(t,t), \quad c = 1, \ldots, C. \tag{A.2}$$

Now let $s \in T \setminus \mathbf{S}$ and $r \in T$ be two other time points such that r, s and t are all distinct. Then $(Y(r), Y(s), Y(t))^T$ follows a mixture of 3-dimensional Gaussian distributions

$$(Y(r), Y(s), Y(t))^T \sim \sum_{c=1}^{C} \lambda_c N_3(\boldsymbol{\nu}_c(r,s,t), \boldsymbol{\Omega}_c(r,s,t)) = \sum_{c=1}^{C} \pi_c N_3(\boldsymbol{\mu}_c(r,s,t), \boldsymbol{\Gamma}_c(r,s,t))$$

where

$$\boldsymbol{\nu}_c(r,s,t) = \begin{pmatrix} \mathbf{x}(r)^T\boldsymbol{\gamma}_c(r) \\ \mathbf{x}(s)^T\boldsymbol{\gamma}_c(s) \\ \mathbf{x}(t)^T\boldsymbol{\gamma}_c(t) \end{pmatrix} \qquad \boldsymbol{\Omega}_c(r,s,t) = \begin{pmatrix} \Omega_c(r,r) & \Omega_c(r,s) & \Omega_c(r,t) \\ \Omega_c(s,r) & \Omega_c(s,s) & \Omega_c(s,t) \\ \Omega_c(t,r) & \Omega_c(t,s) & \Omega_c(t,t) \end{pmatrix}$$

$$\boldsymbol{\mu}_c(r,s,t) = \begin{pmatrix} \mathbf{x}(r)^T\boldsymbol{\beta}_c(r) \\ \mathbf{x}(s)^T\boldsymbol{\beta}_c(s) \\ \mathbf{x}(t)^T\boldsymbol{\beta}_c(t) \end{pmatrix} \qquad \boldsymbol{\Gamma}_c(r,s,t) = \begin{pmatrix} \Gamma_c(r,r) & \Gamma_c(r,s) & \Gamma_c(r,t) \\ \Gamma_c(s,r) & \Gamma_c(s,s) & \Gamma_c(s,t) \\ \Gamma_c(t,r) & \Gamma_c(t,s) & \Gamma_c(t,t) \end{pmatrix}$$

Since $s, t \in T \setminus \mathbf{S}$, from Lemma 2, the distribution of $(Y(r), Y(s), Y(t))^T$ as a mixture of 3-dimensional Gaussian distributions is identifiable, hence the permutation ω_t is actually independent of t. Therefore, for any $t_1, t_2 \in \{r, s, t\}$,

$$\lambda_{\omega(c)} = \pi_c, \qquad \mathbf{X}(t_1)^T\boldsymbol{\gamma}_{\omega(c)}(t_1) = \mathbf{X}(t_1)^T\boldsymbol{\beta}_c(t_1), \qquad \Omega_{\omega(c)}(t_1,t_2) = \Gamma_c(t_1,t_2), \quad c = 1, \ldots, C, \tag{A.3}$$

and for $t_3 \in \{s, t\}$,

$$\boldsymbol{\gamma}_{\omega(c)}(t_3) = \boldsymbol{\beta}_c(t_3), \qquad \Omega_{\omega(c)}(t_3,t_3) = \Gamma_c(t_3,t_3). \tag{A.4}$$

By the continuity of $\boldsymbol{\beta}_c(\cdot)$ and $\Gamma_c(\cdot,\cdot)$, Eq. (A.4) also holds for $t_3 = r$. This completes the proof of identifiability.

Derivation of Estimation procedure 1.

Define the random variables for component membership as

$$z_{ic} = \begin{cases} 1, & \text{if } \{y_i(t), t \in T\} \text{ is in the } c\text{th group,} \\ 0, & \text{otherwise.} \end{cases}$$

The complete likelihood of $\{(y_{ij}, z_{ic}), j = 1, \ldots, N_i, i = 1, \ldots, n, c = 1, \ldots, C\}$ is

$$\prod_{i=1}^{n}\prod_{c=1}^{C}\left[\pi_c \prod_{j=1}^{N_i} \phi\{y_{ij}|X_i(t_{ij})^T\boldsymbol{\beta}_c(t_{ij}), \sigma_c^{*2}(t_{ij})\}\right]^{z_{ic}}.$$

In the E-step, we calculate the expectation of z_{ic} given $\pi_c^{(l)}$, $\sigma_c^{*2(l)}(\cdot)$, and $\boldsymbol{\beta}_c^{(l)}(\cdot)$, $c = 1, \ldots, C$, which is shown in (2.7). In the M-step, we maximize the logarithm of complete log-likelihood function with z_{ic} replaced by $r_{ic}^{(l+1)}$, which is

$$\sum_{i=1}^{n}\sum_{c=1}^{C}\left[r_{ic}^{(l+1)}\log(\pi_c) + r_{ic}^{(l+1)}\sum_{j=1}^{N_i}\log\phi\{y_{ij}|X_i(t_{ij})^T\boldsymbol{\beta}_c(t_{ij}), \sigma_c^{*2}(t_{ij})\}\right].$$

The maximization with respect to π_c leads to (2.8). For nonparametric smoothing functions $\boldsymbol{\beta}_c(\cdot)$ and $\sigma_c^{*2}(\cdot)$, we consider kernel regression for estimation. For any $t_0 \in T$, we approximate $\boldsymbol{\beta}_c(t_{ij})$ by $\boldsymbol{\beta}_c(t_0)$ and $\sigma_c^{*2}(t_{ij})$ by $\sigma_c^{*2}(t_0)$ for t_{ij} in the neighborhood of t_0. Thus, the corresponding local log-likelihood function is

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l+1)}\sum_{j=1}^{N_i}[\log\phi\{y_{ij}|X_i(t_{ij})^T\boldsymbol{\beta}_c(t_0), \sigma_c^{*2}(t_0)\}]K_h(t_{ij} - t_0), \tag{A.5}$$

where $K_h(t)$ is a rescaled kernel function $h^{-1}K(t/h)$ with a kernel function $K(t)$. Maximizing (A.5) with respect to $\boldsymbol{\beta}_c(t_0)$ and $\sigma_c^{*2}(t_0)$ yields (2.9) and (2.10).

References

Aneiros-Pérez, G., Cardot, H., Estévez-Pérez, G., Vieu, P., 2004. Maximum ozone concentration forecasting by functional non-parametric approaches. Environmetrics 15 (7), 675–685.

Aneiros-Pérez, G., Vieu, P., 2006. Semi-functional partial linear regression. Statist. Probab. Lett. 76 (11), 1102–1110.

Bongiorno, E.G., Salinelli, E., Goia, A., Vieu, P., 2014. Contributions in Infinite-Dimensional Statistics and Related Topics. Societa Editrice Esculapio.

Bosq, D., 2000. Linear Processes in Function Spaces. Springer, New York.

Bosq, D., Blanke, D., 2007. Inference and Prediction in Large Dimensions. Wiley.

Cai, Z., Fan, J., Li, R., 2000. Efficient estimation and inferences for varying-coefficient models. J. Amer. Statist. Assoc. 95 (451), 888–902.

Chen, H., Chen, J., Kalbfleisch, J.D., 2001. A modified likelihood ratio test for homogeneity in finite mixture models. J. R. Stat. Soc. Ser. B 63 (1), 19–29.

Chen, D., Hall, P., Müller, H.-G., et al., 2011. Single and multiple index functional regression models with nonparametric link. Ann. Statist. 39 (3), 1720–1747.

DeSarbo, W.S., Cron, W.L., 1988. A maximum likelihood methodology for clusterwise linear regression. J. Classification 5 (2), 249–282.

Ettinger, B., Guillas, S., Lai, M.-J., 2012. Bivariate splines for ozone concentration forecasting. Environmetrics 23 (4), 317–328.

Eubank, R., Huang, C., Maldonado, Y.M., Wang, N., Wang, S., Buchanan, R., 2004. Smoothing spline estimation in varying-coefficient models. J. R. Stat. Soc. Ser. B 66 (3), 653–667.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman, Hall, London.

Fan, J., Zhang, J.-T., 2000. Two-step estimation of functional linear models with applications to longitudinal data. J. R. Stat. Soc. Ser. B 62 (2), 303–322.

Fan, J., Zhang, C., Zhang, J., 1999. Sieve likelihood ratio statistics and wilks phenomenon. Department of Statistics, UCLA.

Fan, J., Zhang, C., Zhang, J., 2001. Generalized likelihood ratio statistics and wilks phenomenon. Ann. Statist. 29 (1), 153–193.

Ferraty, F., Goia, A., Salinelli, E., Vieu, P., 2013. Functional projection pursuit regression. Test 22 (2), 293–320.

Ferraty, F., Romain, Y., 2011. The Oxford Handbook of Functional Data Analaysis. Oxford University Press.

Ferraty, F., Van Keilegom, I., Vieu, P., 2012. Regression when both response and predictor are functions. J. Multivariate Anal. 109, 10–28.

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer.

Garnaut, R., Howes, S., Jotzo, F., Sheehan, P., 2008. Emissions in the platinum age: the implications of rapid development for climate-change mitigation. Oxf. Rev. Econ. Policy 24 (2), 377–401.

Hartigan, J., 1985. A failure of likelihood asymptotics for normal mixtures. In: Le Cam, L., Olshen, R.A. (Eds.), Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, 2. Wadsworth, Belmont, CA, pp. 807–810.

Heard, N., Holmes, C., Stephens, D., 2006. A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes. J. Amer. Statist. Assoc. 101 (473), 18–29.

Hennig, C., 2000. Identifiablity of models for clusterwise linear regression. J. Classification 17 (2), 273–296.

Hoover, D.R., Rice, J.A., Wu, C.O., Yang, L.-P., 1998. Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. Biometrika 85 (4), 809–822.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications, Vol. 200. Springer Science & Business Media.

Hoshikawa, T., 2013. Mixture regression for observational data, with application to functional regression models. http://arxiv.org/abs/1307.0170.

Huang, M., Li, R., Wang, H., Yao, W., 2014. Estimating mixture of gaussian processes by kernel smoothing. J. Bus. Econom. Statist. 32 (2), 259–270.

Huang, M., Yao, W., 2012. Mixture of regression models with varying mixing proportions: a semiparametric approach. J. Amer. Statist. Assoc. 107 (498), 711–724.

James, G., Sugar, C., 2003. Clustering for sparsely sampled functional data. J. Amer. Statist. Assoc. 98 (462), 397–408.

Kudraszow, N.L., Vieu, P., 2013. Uniform consistency of knn regressors for functional variables. Statist. Probab. Lett. 83 (8), 1863–1870.

Leroux, B.G., 1992. Consistent estimation of a mixing distribution. Ann. Statist. 20 (3), 1350–1360.

Li, P., Chen, J., 2010. Testing the order of a finite mixture. J. Amer. Statist. Assoc. 105 (491).

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 13–22.

Lin, X., Carroll, R., 2000. Nonparametric function estimation for clustered data when the predictor is measured without/with error. J. Amer. Statist. Assoc. 520–534.
Lu, Z., Song, X., 2012. Finite mixture varying coefficient models for analyzing longitudinal heterogenous data. Stat. Med. 31 (6), 544–560.
Luan, Y., Li, H., 2003. Clustering of time-course gene expression data using a mixed-effects model with b-splines. Bioinformatics 19 (4), 474–482.
Ma, P., Zhong, W., 2008. Penalized clustering of large-scale functional data with multiple covariates. J. Amer. Statist. Assoc. 103 (482), 625–636.
Marron, J.S., 1988. Automatic smoothing parameter selection: a survey. Empir. Econ. 13 (3–4), 187–208.
McLachlan, G., Peel, D., 2000. Finite Mixture Models, Vol. 299. John Wiley & Sons.
Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis, Vol. 77. Springer, New York.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer, New York.
Rice, J., Silverman, B., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. J. R. Stat. Soc. Ser. B 233–243.
Sapatnekar, S.S., 2011. Overcoming variations in nanometer-scale technologies. IEEE J. Emerg. Sel. Top. Circuits Syst. 1 (1), 5–18.
Titterington, D., Smith, A., Makov, U., et al., 1985. Statistical Analysis of Finite Mixture Distributions. Wiley, New York.
Verbeke, G., Lesaffre, E., 1996. A linear mixed-effects model with heterogeneity in the random-effects population. J. Amer. Statist. Assoc. 91 (433), 217–221.
Xu, W., Hedeker, D., 2001. A random-effects mixture model for classifying treatment response in longitudinal clinical trials. J. Biopharm. Statist. 11 (4), 253–273.
Yakowitz, S., Spragins, J., 1968. On the identifiability of finite mixtures. Ann. Math. Stat. 39 (1), 209–214.
Yao, F., Fu, Y., Lee, T.C., 2011. Functional mixture regression. Biostatistics 12 (2), 341–353.
Yao, F., Müller, H., Wang, J., 2005. Functional data analysis for sparse longitudinal data. J. Amer. Statist. Assoc. 100 (470), 577–590.
Young, D., Hunter, D., Elmore, R., Xuan, F., Hettmansperger, T., Thomas, H., 2007. The mixtools package: tools for mixture models. R Package Version 0.2.0.