Estimating Mixture of Gaussian Processes by Kernel Smoothing

Mian Huang, Runze Li, Hansheng Wang, and Weixin Yao

October 20, 2013

Abstract

When the functional data is not homogeneous, e.g., there exist multiple classes of functional curves in the dataset, traditional estimation methods may fail. In this paper, we propose a new estimation procedure for the Mixture of Gaussian Processes, to incorporate both functional and inhomogeneous properties of the data. Our method can be viewed as a natural extension of high-dimensional normal mixtures. However, the key difference is that smoothed structures are imposed for both the mean and covariance functions. The model is shown to be identifiable, and can be estimated efficiently by a novel combination of the ideas from EM algorithm, kernel regression, and functional principal component analysis. Our methodology is empirically justified by Monte Carlo simulations and illustrated by an analysis of a supermarket dataset.

Keywords: Identifiability, EM algorithm, Kernel regression, Gaussian process, Functional principal component analysis

[1]Mian Huang is Associate Professor, School of Statistics and Management and Key Laboratory of Mathematical Economics at SHUFE, Ministry of Education, Shanghai University of Finance and Economics (SHUFE), Shanghai, 200433, P. R. China. Email:huang.mian@shufe.edu.cn. Runze Li is Distinguished Professor, Department of Statistics and The Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111. Email: rzli@psu.edu. Hansheng Wang is Professor, Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University, Beijing, 100871, P. R. China. Email: hansheng@gsm.pku.edu.cn. Weixin Yao is Associate Professor, Department of Statistics, Kansas State University, Manhattan, Kansas 66506. Email: wxyao@ksu.edu. Huang's research is supported by National Natural Science Foundation of China (NNSFC), Grant 11301324. Li's research was supported by National Institute on Drug Abuse (NIDA) grants R21 DA024260 and P50-DA10075, National Cancer Institute grant R01 CA168676 and NNSFC grant 11028103. Wang's research was supported in part by NSFC grants 11131002, 11271032, Fox Ying Tong Education Foundation, the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University. The authors thank the Editor, the AE and reviewers for their constructive comments, which have led to a dramatic improvement of the earlier version of this paper. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NIDA.

1 Introduction

The rapid development of information technologies enables researchers to collect and store functional data at a low cost. As a result, the quantitative analysis of functional data becomes practically feasible; see Ramsay and Silverman (2005) for a comprehensive and excellent treatment. The basis of functional data analysis consists of the estimations of the mean function and the covariance structure. Among many approaches, functional principal component (FPC) analysis serves as a key technique in functional data analysis. Rice and Silverman (1991) and James et al. (2000) studied the spline smoothing methods in FPC analysis; Staniswalis and Lee (1998) and Yao et al. (2003) applied kernel-based smoothing methods for FPC analysis in irregular and sparse longitudinal data. The asymptotic properties of principal component functions are investigated in Yao et al. (2005) and Hall et al. (2006).

For an illustration of functional data, Figure 1 depicts the plot of a set of collected curves. This dataset contains the number of customers who visited a particular supermarket in China on each of 139 days. For each day, the number of customers shopping in the supermarket is observed every half hour from 7:00am to 5:30pm. Thus, there are 22 observations for each day. The collected time was coded as 1 for 7:00am, 2 for 7:30am, and so on. In the analysis of this dataset, we regard each day as one subject. Thus, we have a total of 139 subjects. Figure 1 shows that the variability may be large in certain time periods. Intuitively, the customer flow (i.e., the number of customers) may show different patterns in weekdays, weekends and holiday season, and hence the data are likely inhomogeneous. Although the nominal identity (weekday, weekend, or holiday) of a subject is known, they may switch to form a long-holidays by national or local government policies, e.g., the holiday week of national day. In this paper, we will treat the identities as unknown. To statistically model such inhomogeneity for the multivariate response, we may simply consider a mixture of 22-dimensional multivariate normal distributions. Nevertheless, we find this method less effective because the 22×22 covariance matrices for each component have to be estimated. This has been an inevitable step for a general normal mixture model. With such a limited sample size (i.e, 139), the estimated covariance matrices are likely to be ill-conditioned. As a consequence, the estimation accuracy of its inverse is very poor. In addition, if the data are collected at irregular time points, the covariance structure will be different for different subjects and thus the mixture of multivariate normal distribution cannot be applied, even when the sample size is large. This

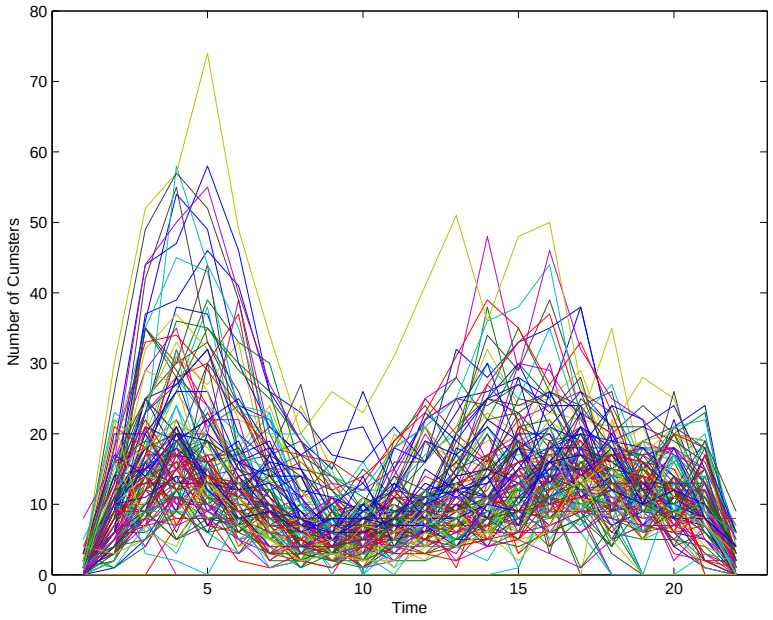


Figure 1: Plot of supermarket data.

motivates us to develop new methods for analysis of inhomogeneous functional data.

Mixture of Gaussian processes is an interesting and useful alternative to mixture of high-dimensional normals. In this paper, we propose a new smooth estimation procedure for mixture of Gaussian processes. Compared with a general normal mixture, the major advantage of our method is that smoothed structures are imposed for both the mean and covariance functions. Within this new framework, the unknown functions can be estimated efficiently by a novel combination of the ideas from EM algorithm, kernel regression, and functional principal component analysis. Therefore, the challenging task of high-dimensional covariance matrix estimation can be completely avoided. In addition, the proposed mixture models can deal with data collected at irregular, possibly subject depending time points. It is clear that a mixture of multivariate normals is not applicable for such data.

James and Sugar (2003) considered a general functional model for clustering functional data, which is indeed a mixture of Gaussian processes. In their approach, they represented individual curves by natural cubic splines, and imposed some parametric assumptions and restrictions on the spline coefficients. This version of the mixture of Gaussian processes is casted as a structural parametric finite mixture of normals, which is referred to as the functional clustering model.

Maximum likelihood and EM algorithm are developed for model estimation. Functional clustering models have been studied and applied in literature. In genetic research, Luan and Li (2003) considered a functional clustering model for time-course gene expression data, in which B-spline are used to model the mean and covariance function of each component. Bayesian approaches for functional clustering models are studied in Heard et al. (2006), and Ma and Zhong (2008).

In this paper, we shall systematically study the mixture of Gaussian processes. We first prove that the the mixture of Gaussian processes is identifiable under mild conditions. We propose new estimation procedures using kernel regression and modified EM-type algorithms. We introduce functional principal component analysis for the estimation procedure, which provides the advantage of effective computation, e.g., avoids the inverse of high-dimensional covariance matrix, and facilitates the covariance estimation. Functional principal component analysis also provides a powerful tool to interpret the results via the eigenvalue and eigenfunctions. Practical guides for model selection are addressed, and a bootstrap procedure for constructing confidence intervals is proposed. We empirically justify these estimation procedures by Monte Carlo simulations, and an illustration in real data analysis, including a detailed interpretation of the estimated functional principal components.

The rest of this paper is structured as follows. We present the definition of mixture of Gaussian processes and give the identifiability result in Section 2. In Section 3, we develop estimation procedures for the newly proposed models. Simulation results and an empirical analysis of supermarket dataset are presented in Section 4. Concluding remarks and some discussions are given in Section 5. Proof is given in the appendix.

2 Model and Identifiability

Let $\mathcal{C}$ be a latent class variable with a discrete distribution $P(\mathcal{C} = c) = \rho_c$ for $c = 1, 2, \cdots, C$. It is assumed in this paper that C is fixed and known. We will briefly discuss how to determine C in Section 3. Given $\mathcal{C} = c$, $\{X(t), t \in \mathbb{T}\}$ follows a Gaussian process with mean $\mu_c(t)$ and covariance function $\mathrm{Cov}\{X(s), X(t)\} = G_c(s,t)$. We refer to $\{X(t) : t \in \mathbb{T}\}$ as a mixture of Gaussian processes. Typically, $\mathbb{T}$ is a closed and bounded time interval $[0, T]$. It is assumed throughout this paper that $\mu_c(t)$ is a smooth function of t, and $G_c(s,t)$ is a positive definite and bivariate smooth function of s and t. Thus, the path of $X(t)$ indeed is a smooth function.

We first study the identifiability of the proposed mixture of Gaussian processes (Proof is given in the Appendix).

Theorem 1 *Suppose $G_c(s,t)$ is a positive definite and bivariate smooth function of s and t and $\mu_c(t)$ is a smooth function of t for any $c = 1, \ldots, C$. Let $\mathbf{S} = \{t \in \mathbb{T} : (\mu_i(t), G_i(t,t)) = (\mu_j(t), G_j(t,t))$ for some $1 \leq i \neq j \leq C\}$. If the complement of $\mathbf{S}$ is not empty, then the above proposed mixture of Gaussian processes is identifiable.*

The covariance function $G_c(s,t)$ can be represented as

$$G_c(s,t) = \sum_{q=1}^{\infty} \lambda_{qc} v_{qc}(t) v_{qc}(s),$$

where λ_{qc}'s are eigenvalues, and $v_{qc}(\cdot)$'s are eigenfunctions. Furthermore, we have $\lambda_{1c} \geq \lambda_{2c} \geq \cdots$, and $\sum_q \lambda_{qc} < \infty$, for $c = 1, \cdots, C$. By the Karhunen-Loève theorem, if the i-th subject $X_i(t)$ is from the c-th component, then it can be represented as follows

$$X_i(t) = \mu_c(t) + \sum_{q=1}^{\infty} \xi_{iqc} v_{qc}(t),$$

where the functional principal component score ξ_{iqc} is considered as independent random variables with $\mathrm{E}(\xi_{iqc}) = 0$, and $\mathrm{Var}(\xi_{iqc}) = \lambda_{qc}$.

Since the sample path of $X_i(t)$ is a smooth function of t, $X_i(t)$ is termed a smooth random function (Yao et al., 2005) . As depicted in Figure 1, the collected sample of random curves are typically not smooth in practice. Following Yao et al. (2003), it is assumed that the observed curve $\{y_i(t), t = t_{ij}, j = 1, \cdots, N_i\}$ is

$$y_i(t) = X_i(t) + \epsilon_i(t),$$

where $\epsilon_i(t)$ is additive measurement error, and it is assumed that $\epsilon_i(t_{ij})$, for all i and j, are independent and identically distributed as $N(0, \sigma^2)$. Denote $y_{ij} = y_i(t_{ij})$ and $\epsilon_{ij} = \epsilon_i(t_{ij})$. Throughout this paper, it is assumed that conditioning on $\mathcal{C} = c$, the observations y_{ij}, $j = 1, \cdots, N_i$ and $i = 1, \cdots, n$, follows

$$y_{ij} = \mu_c(t_{ij}) + \sum_{q=1}^{\infty} \xi_{iqc} v_{qc}(t_{ij}) + \epsilon_{ij}, \tag{2.1}$$

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

We also consider a reduced model from model (2.1), where the data within subjects are independent. This means that $G_c(s,t) = 0$ if $s \neq t$. Let $\sigma_c^{*2}(t) = G_c(t,t) + \sigma^2$, it follows that conditioning on $\mathcal{C} = c$

$$y_{ij} = \mu_c(t_{ij}) + \epsilon_{ij}^*, \tag{2.2}$$

where ϵ_{ij}^* are independent with $E(\epsilon_{ij}^*) = 0$ and $\text{Var}(\epsilon_{ij}^*) = \sigma_c^{*2}(t_{ij})$. This is equivalent to treating y_{ij}s sampled from the following distribution:

$$y(t) \sim \sum_{c=1}^{C} \rho_c N\{\mu_c(t), \sigma_c^{*2}(t)\}. \tag{2.3}$$

Theorem 2 *Suppose $\mu_c(t)$ and $\sigma_c^*(t)$ are smooth functions of t for any $c = 1, \ldots, C$. Let $\mathbf{S}^* = \{t \in \mathbb{T} : (\mu_i(t), \sigma_i^*(t)) = (\mu_j(t), \sigma_i^*(t))$ for some $1 \leq i \neq j \leq C\}$. If the complement of $\mathbf{S}^*$ is not empty, then the mixture model (2.3) is identifiable.*

The proof of Theorem 2 (omitted) is similar to Theorem 1.

3 Estimation Procedures

3.1 Estimation of Model (2.3)

Denote by $\phi(y|\mu, \sigma^2)$ the density function of $N(\mu, \sigma^2)$. Then for model (2.3), the log-likelihood function of the collected data is

$$\sum_{i=1}^{n} \log \left[\sum_{c=1}^{C} \rho_c \prod_{j=1}^{N_i} \phi \left\{ y_{ij} | \mu_c(t_{ij}), \sigma_c^{*2}(t_{ij}) \right\} \right]. \tag{3.1}$$

We now propose an EM-type algorithm to maximize (3.1). Define the membership identity random variables

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

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

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

After the l-th iteration of the EM algorithm, suppose that we have $\rho_c^{(l)}$, $\sigma_c^{*2(l)}(\cdot)$, and $\mu_c^{(l)}(\cdot)$. Thus, in the E-step of the $(l+1)$-th iteration, the expectation of the latent variable z_{ic} is given by

$$r_{ic}^{(l+1)} = \frac{\rho_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij}|\mu_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\}\right]}{\sum_{c=1}^{C} \rho_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_{ij}|\mu_c^{(l)}(t_{ij}), \sigma_c^{*2(l)}(t_{ij})\}\right]}. \tag{3.2}$$

In the M-step of the $(l+1)$-th iteration, we would maximize the logarithm of complete log-likelihood function with z_{ic} replaced by $r_{ic}^{(l+1)}$, which is

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

This leads to

$$\rho_c^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l+1)}. \tag{3.3}$$

Note that both $\mu_c(\cdot)$ and $\sigma_c^{*2}(\cdot)$ are nonparametric smoothing functions. Here we use kernel regression to estimate $\mu_c(\cdot)$'s and $\sigma_c^{*2}(\cdot)$'s. For any $t_0 \in T$, we approximate $\mu_c(t_{ij})$ by $\mu_c(t_0)$ 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

$$\sum_{i=1}^{n}\sum_{c=1}^{C} r_{ic}^{(l+1)}\sum_{j=1}^{N_i}[\log\phi\{y_{ij}|\mu_c(t_0), \sigma_c^{*2}(t_0)\}]K_h(t_{ij}-t_0), \tag{3.4}$$

where $K_h(t)$ is a rescaled kernel function $h^{-1}K(t/h)$ with a kernel function $K(t)$. Maximizing (3.4) with respect to $\mu_c(t_0)$ and $\sigma_c^{*2}(t_0)$, $c = 1, \cdots, C$, yields

$$\mu_c^{(l+1)}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ijc}^{(l+1)} y_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ijc}^{(l+1)}}, \tag{3.5}$$

$$\sigma_c^{*2(l+1)}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ijc}^{(l+1)}\{y_{ij} - \mu_c^{(l+1)}(t_0)\}^2}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ijc}^{(l+1)}}, \tag{3.6}$$

where $w_{ijc}^{(l+1)} = r_{ic}^{(l+1)}K_h(t_{ij}-t_0)$. In practice, we evaluate the estimates at a set of grid points for the given label in the E-step. Let $\{u_1, \cdots, u_{n_{grid}}\}$ be a set of grid points at which the estimated functions are evaluated, where n_{grid} is the number of grid points. If the total number of observations $J = \sum_{i=1}^{n} N_i$, is not very large, we can directly use all the time points as the grid points. Otherwise, we update $\mu_c(t_{ij})$ and $\sigma_c^{*2}(t_{ij})$, $i = 1, \cdots, n, j = 1, \cdots, N_i$ by linearly interpolating $\mu_c^{(l+1)}(u_k)$ and $\sigma_c^{*2(l)}(u_k)$, $k = 1, \cdots, n_{grid}$. Denote by $\tilde{\rho}_c$, $\tilde{\mu}_c(\cdot)$, and $\tilde{\sigma}_c^{*2}(\cdot)$ the resulting estimate of ρ_c, $\mu_c(\cdot)$, and $\sigma_c^{*2}(\cdot)$, respectively.

3.2 Estimation of Model (2.1)

3.2.1 Initial Estimation

For a Gaussian process, it is inevitable to estimate the mean functions first, and then estimate the covariance function based on the residuals. As demonstrated in Lin and Carroll (2000), the kernel generalized estimating equation (GEE) method for repeated measurement data yields an optimal estimate in a certain sense by pretending the data within subjects are independent. Furthermore, kernel GEE method with working independent covariance structure is easy to implement. Therefore for the mixture of Gaussian processes, it is natural to adapt the estimation procedure of model (2.3), and pretending that the data within subjects are independent. We refer to this procedure as an initial estimation with working independent correlation. This yields the initial estimation of the mean functions and probability identities of each subject.

3.2.2 Estimation of Covariances

We now deal with estimation of covariance functions using functional principal analysis. Let $\bar{G}_{ic}(t_{ij}, t_{il}) = \{y_{ij} - \tilde{\mu}_c(t_{ij})\}\{y_{il} - \tilde{\mu}_c(t_{il})\}$. Note that given $\mathcal{C} = c$, $\mathrm{Cov}\{Y(t), Y(t)\} = G_c(t,t) + \sigma^2$, and $\mathrm{Cov}\{Y(s), Y(t)\} = G_c(s,t)$ for $s \neq t$. If z_{ic} were observable, then the covariance function $G_c(s,t)$ could be estimated by a two-dimensional kernel smoother, which is to minimize

$$\sum_{i=1}^{n} z_{ic} \sum_{1 \leq j \neq l \leq N} [\bar{G}_{ic}(t_{ij}, t_{il}) - \beta_0]^2 K_{h^*}(t_{ij} - s) K_{h^*}(t_{il} - t), \tag{3.7}$$

with respect to β_0. In practice, z_{ic} is a latent variable. Following the idea of the EM algorithm, we replace z_{ic} by its expectation r_{ic} given in (3.2), which was obtained in the initial estimation procedure with working independent correlation. Thus, we minimize

$$\sum_{i=1}^{n} r_{ic} \sum_{1 \leq j \neq l \leq N} [\bar{G}_{ic}(t_{ij}, t_{il}) - \beta_0]^2 K_{h^*}(t_{ij} - s) K_{h^*}(t_{il} - t), \tag{3.8}$$

with respect to β_0. The minimizer $\hat{G}_c(s,t) \equiv \hat{\beta}_0$ of (3.8) has a closed form solution, given by

$$\hat{G}_c(s,t) = \frac{\sum_{i=1}^{n} r_{ic} \sum\limits_{1 \leq j \neq l \leq N_i} \bar{G}_{ic}(t_{ij}, t_{il}) K_{h^*}(t_{ij} - s) K_{h^*}(t_{il} - t)}{\sum\limits_{i=1}^{n} r_{ic} \sum\limits_{1 \leq j \neq l \leq N_i} K_{h^*}(t_{ij} - s) K_{h^*}(t_{il} - t)}. \tag{3.9}$$

Following Rice and Silverman (1991), the estimation of eigenvalues and eigenfunctions are based on discretizing the covariance estimate $\hat{G}_c(s,t)$. The estimates of eigenvalues $\hat{\lambda}_{qc}$ and eigenfunctions $\hat{v}_{qc}(\cdot)$ are determined by eigenfunctions

$$\int_T \hat{G}_c(s,t)\hat{v}_{qc}(s)ds = \hat{\lambda}_{qc}\hat{v}_{qc}(t), \tag{3.10}$$

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 $p \neq q$. Then, in order for the resulting estimate of $G_c(s,t)$ to be positive definite, we set

$$\hat{G}_c(s,t) = \sum_q \hat{\lambda}_{qc} I(\hat{\lambda}_{qc} > 0)\hat{v}_{qc}(s)\hat{v}_{qc}(t).$$

3.2.3 An Iterative Estimation Procedure

Given $\hat{\mu}_c(t)$ and $\hat{v}_{qc}(t)$, the functional principal component score ξ_{iqc} can be estimated by

$$\hat{\xi}_{iqc} = \int_T \{y_i(t) - \hat{\mu}_c(t)\}\, \hat{v}_{qc}(t)dt. \tag{3.11}$$

Furthermore, for $j = 1, \cdots, N_i$ and $i = 1, \cdots, n$, define

$$\hat{\eta}_{ic}(t_{ij}) = \sum_q \hat{\xi}_{iqc} I(\hat{\lambda}_{qc} > 0)\hat{v}_{qc}(t_{ij}), \tag{3.12}$$

which is an estimate of $\eta_{ic}(t_{ij}) = \sum_q \xi_{iqc} I(\lambda_{qc} > 0)v_{qc}(t_{ij})$. Let

$$y_c^*(t_{ij}) = y_{ij} - \hat{\eta}_{ic}(t_{ij}). \tag{3.13}$$

Then, conditioning on $\mathcal{C} = c$, model (2.1) can be approximated by

$$y_c^*(t_{ij}) \approx \mu_c(t_{ij}) + \epsilon_{ij}, \tag{3.14}$$

where ϵ_{ij}'s are independent and identically distributed as $N(0, \sigma^2)$. Hence, with the aid of functional PCA, we can transform the correlated data to uncorrelated data with a few eigenvalues and eigenfunctions from the estimate of $G_c(s,t)$. Based on $\{y_c^*(t_{ij}), i = 1, \ldots, n, j = 1, \ldots, N_i, c = 1, \ldots, C\}$, the EM-type algorithm for model (2.2) can be adapted to further improve the estimate of $\mu_c(t)$, σ^2, and ρ_cs. Slight revision is made according to the constant variance of (3.14), which is different from (2.2). Specifically, in the E-step we find the probability

$$r_{ic}^{(l+1)} = \frac{\rho_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_c^*(t_{ij})|\mu_c^{(l)}(t_{ij}), \sigma^{2(l)}\}\right]}{\sum_{c=1}^{C} \rho_c^{(l)} \left[\prod_{j=1}^{N_i} \phi\{y_c^*(t_{ij})|\mu_c^{(l)}(t_{ij}), \sigma^{2(l)}\}\right]}. \tag{3.15}$$

In the M-step, we update the estimates of $\mu_c(t)$, ρ_c, and σ^2. For $t_0 \in \{u_1, \cdots, u_{n_{grid}}\}$,

$$\mu_c^{(l+1)}(t_0) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ijc}^{(l+1)} y_c^*(t_{ij})}{\sum_{i=1}^{n}\sum_{j=1}^{N_i} w_{ijc}^{(l+1)}}, \tag{3.16}$$

where $w_{ijc}^{(l+1)} = r_{ic}^{(l+1)} K_h(t_{ij} - t_0)$, and

$$\rho_c^{(l+1)} = \frac{1}{n}\sum_{i=1}^{n} r_{ic}^{(l+1)}, \tag{3.17}$$

$$\sigma^{2(l+1)} = \frac{1}{\sum_{i=1}^{n} N_i}\sum_{i=1}^{n}\sum_{c=1}^{C}\sum_{j=1}^{N_i} r_{ic}^{(l+1)}\{y_{ij} - \mu_c^{(l+1)}(t_{ij})\}^2. \tag{3.18}$$

Furthermore, we update $\{\mu_c^{(l+1)}(t_{ij}), i = 1, \cdots, n, j = 1, \cdots, N_i\}$ by linearly interpolating $\mu_c^{(l+1)}(u_k)$, $k = 1, \cdots, n_{grid}$.

To improve the estimation, we further propose an iterative estimation procedure, which iterates between one cycle of the above procedure, and the estimation of the covariance structure. The proposed estimation procedure can be summarized as follows:

An Iterative Estimation Procedure

Step 1: Calculate $\tilde{\mu}_c(\cdot)$ using the EM-type algorithm of (3.2)–(3.6).

Step 2: Given $\mu_c(\cdot)$, and r_{ic}s, obtain $\hat{G}_c(s,t)$ using (3.9) and calculate $\hat{\eta}_{ic}(t_{ij})$ using (3.10), (3.11), and (3.12).

Step 3: Calculate $y_c^*(t_{ij})$ in (3.13), update $\mu_c(t), \sigma^2$, ρ_c, and r_{ic}s using (3.15)–(3.18).

Iteratively calculate Step 2 and Step 3 until convergence. It is worth noting that this procedure is easy to implement, since it avoids the disadvantages of high-dimensional mixture of normals, i.e., the calculation of inverse of the covariance matrix.

Remark. For model (2.1), when the components are well separated, the initial estimation procedure estimates the mean functions almost as well as the iterative procedure which incorporates the correlations. When the components are very separated, the component identities of the samples can be considered as known. Therefore, the problem is similar to the traditional homogenous functional data analysis. Typically, the estimated covariance has a slower convergence rate than the estimated mean function, and the convergence rate of the eigenfunctions relates to rate of

estimated covariance (Yao et al., 2005). Hence, estimating mean function by incorporating correlation via the estimated eigenfunctions can not be more efficient. However, when the components are overlapped, estimation with incorporating correlation can improve the estimation of component identities, and therefore improve both estimations of mean and covariance functions of each component. We will design simulation study to illustrate this point in Section 4.

3.3 Practical Implementation Issues

Now we address some important practical issues, including the choice of the number of components, bandwidth, and number of eigenfunctions. In practice they may be determined in the following sequence. The number of components shall be determined before the bandwidth and number of eigenfunctions. Once we choose number of components, we select the bandwidths for model (2.2) and the covariance estimates. With the selected bandwidths, we then choose the number of eigenfunctions for each component. Finally we select the bandwidths for the refined estimation procedure for mean functions, and the iterative estimation procedure.

Choice of the number of components. Choosing the number of components C is a critical issue for mixture models. This paper assumes the number of components is known. But when the observations are dense, we may use a simple approach to determine C by using the information criteria for finite mixture of low dimensional multivariate normals. Direct implementation of the information criteria for mixture of Gaussian processes is difficult since the degrees of freedom for mixture of Gaussian processes is not well defined. As a practical alternative, we recommend applying the AIC or BIC with a finite mixture of multivariate normals for part of the observed data. Specifically, for the supermarket data introduced in Section 1, if the data are observed at $(t_1, \cdots, t_N)$ for all subjects, then we may take the partial data observed at $(t_{k_1}, \cdots, t_{k_{N'}})$, a subset of $(t_1, \cdots, t_N)$. In practice, the subsect $(t_{k_1}, \cdots, t_{k_{N'}})$ can be every d points of $(t_1, \cdots, t_N)$ for some $d \geq 2$. For irregular and unbalanced data, one may either bin the data over the observed times or interpolate the data over a regular grid points, and then further use the AIC or BIC to the selected part of the binned data or interpolated data. By using partial data, we are able to determine C before analysis using the proposed procedure, and avoid the disadvantages of high-dimensional mixtures of normals. This has been implemented in the real data analysis in Section 4.2. For sparse data, further research is needed.

Bandwidth selection. Bandwidth selection is another important issue to be addressed. For initial estimation based on model (2.2), we use the same bandwidth for mean and variance functions for simplicity of computation, and the optimal bandwidth can be determined via multi-fold cross-validation (CV) method. For the covariance functions in Section 3.2.2, we may use one-curve-leave-out cross-validation to choose this smoothing parameter, which has been suggested in the literature of covariance function smoothing (Rice and Silverman, 1991; Yao et al., 2005). We also consider the generalized cross-validation (GCV) method given by the released codes associated with Yao et al. (2005). The bandwidth selection in the refined estimation in Section 3.2.3 only involves the mean function, and it can be determined by CV or GCV method. The simulation results in Section 4 demonstrate that the proposed estimation procedure works quite well in a wide range of bandwidths.

Choice of the number of eigenfunctions. A proper number of eigenfunctions is vital to provide a reasonable approximation to the Gaussian process in each component. Rice and Silverman (1991) suggested using the cross-validation method based on the one-curve-leave-out prediction error. Yao et al. (2005) investigated AIC-type criteria in functional principal component analysis, and found that while the AIC and cross-validation give similar results, the AIC is computationally more efficient than cross-validation method. In practice, empirical criteria are also useful to select the number of eigenfunctions. We may choose the number of eigenfunctions so that the percentage of total variation explained by the eigenfunctions is above a certain threshold, e.g., 85 percent or 90 percent.

4 Simulation and Application

In this section, we conduct numerical simulations to demonstrate the performance of the proposed estimation procedures. To assess the performance of the estimates of the unknown regression functions $\mu_c(t)$, we consider the square root of the average squared errors (RASE) for mean functions,

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

where $\{u_j, j = 1, \cdots, n_{grid}\}$ are the grid points at which the unknown functions $\mu_c(\cdot)$ are evaluated. For simplification, the grid points are taken evenly on the range of the t_{ij}s. In the simulation, we

set $n_{grid} = 50$. Similarly, we can define the RASE of the eigenfunctions for the c-th component, which is

$$\text{RASE}_{v_c}^2 = n_{grid}^{-1} \sum_{q=1}^{Q_c} \sum_{j=1}^{n_{grid}} \{\hat{v}_{qc}(u_j) - v_{qc}(u_j)\}^2.$$

where Q_c is the number of eigenfunctions chosen as discussed in Section 3.4. We are also interested in the average of mean square of predicted error, given by

$$\text{MSE} = (\sum_{i=1}^{n} N_i)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{N_i} \left\{ y_{ij} - \sum_{c=1}^{C} \hat{r}_{ic} \hat{X}_{ic}(t_{ij}) \right\}^2,$$

where $\hat{X}_{ic}(t_{ij}) = \hat{\mu}_c(t_{ij}) + \hat{\eta}_c(t_{ij})$. MSE can be considered as a natural estimate of σ^2.

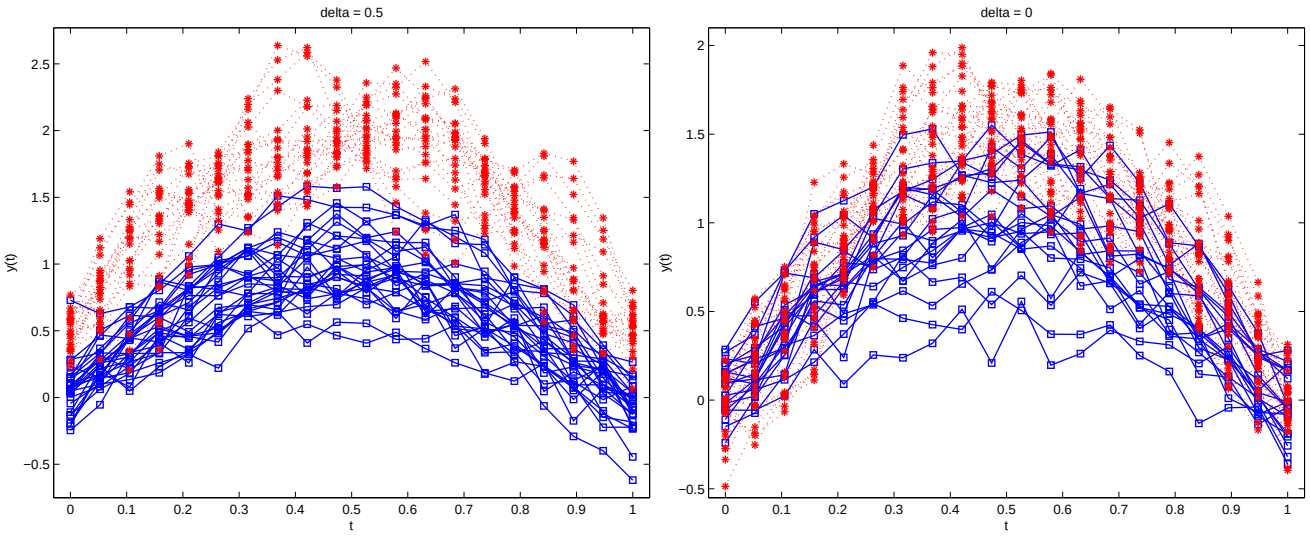
For confidence intervals and standard errors, we consider a bootstrap procedure. Given the observed time $\{t_{ij}, j = 1, \cdots, N_i\}$, we generate a multivariate normal bootstrap sample $\{y^b(t_{ij}), j = 1, \cdots, N_i\}$ with probability $\hat{\rho}_c$, where $\text{E}y^b(t) = \hat{\mu}_c(t)$, and $\text{Cov}(y^b(t), y^b(s)) = \hat{G}_c(t, s) + \hat{\sigma}^2 I$. Then we obtain the standard errors and confidence intervals by using our estimation procedures in each of the bootstrapped samples.

4.1 Simulation Study

In the following example, we generate data from a two-component mixture of Gaussian processes with

$$\begin{aligned}
&\rho_1 = 0.45, \quad \rho_2 = 1 - \rho_1 = 0.55, \quad \text{and} \quad \sigma^2 = 0.01, \\
&\mu_1(t) = \sin(\pi t), \text{ and } \mu_2(t) = \delta + 1.5 \sin(\pi t), \\
&v_{11}(t) = \sqrt{2} \sin(\pi t), \text{ and } v_{12}(t) = \sqrt{2} \cos(\pi t), \\
&v_{21}(t) = \sqrt{2} \sin(4\pi t), \text{ and } v_{22}(t) = \sqrt{2} \cos(4\pi t).
\end{aligned}$$

The simulated data with sample size $n = 100$ are observed at grid points $\{k/N, k = 1, \cdots, N\}$ for both components, where N is set to be 20 and 40. Note that in this example, the data are balanced. However, the computation will be similar for unbalanced data. Let the eigenvalues for both components be $\lambda_{11} = 0.04$, $\lambda_{12} = 0.01$, $\lambda_{21} = 0.04$, $\lambda_{22} = 0.01$, and $\lambda_{qc} = 0$, for $q > 2$, $c = 1, 2$, and let the principal component scores ξ_{iqc} be generated from $N(0, \lambda_{qc})$, $q = 1, 2$, and $c = 1, 2$.



(a) (b)

Figure 2: (a) Typical sample data for the well-separated setting, $\delta = 0.5$; (b) Typical sample data for the heavy-overlap setting $\delta = 0$.

We consider two scenarios of simulation data sets from the above data generation scheme. In the first scenario, we set $\delta = 0.5$. As demonstrated in the typical sample depicted in Figure 2 (a), the subjects from the two components are well separated for this scenario. In the second scenario, we set $\delta = 0$, and the mean functions of the two components are close to each other. Thus, the subjects from the two components are heavily overlapping. A typical sample generated from this scenario is depicted in Figure 2 (b). We compare the performance of two estimation procedures: the estimation of (2.3) using the EM-type algorithm, referred to as procedure of 'working independent'; and the estimation of (2.1) using the iterative estimation procedure, referred to as procedure of 'incorporating correlation'. The comparisons are conducted in both the well-separated setting, and the heavy-overlap setting. For the heavy-overlap setting, we further investigate the performance of eigenfunction estimation.

In the simulation, we assume that the number of components C is known, and use the Epanechnikov kernel for functional smoothing. The bandwidths of mean functions and covariance functions are obtained by CV methods. In simulation we used a fixed bandwidth pair $(\hat{h}_\mu, \hat{h}_{cov})$ for each simulated data. This pair was selected as the average of optimal CV bandwidths of several simulated dataset. Our experience shows that for a wide range of $\hat{h}_{cov}$ including the optimum one,

Table 1: Estimation of Mean functions and ρ_1

		Working independent		Incorporating correlation	
N	δ	RASE_μ	$\rho_1 = 0.45$	RASE_μ	$\rho_1 = 0.45$
20	0.5	0.059(0.012)	0.441(0.049)	0.058(0.012)	0.448(0.049)
	0	0.128(0.035)	0.301(0.048)	0.059(0.012)	0.465(0.050)
40	0.5	0.053(0.014)	0.443(0.047)	0.052(0.014)	0.450(0.047)
	0	0.113(0.031)	0.317(0.048)	0.052(0.014)	0.457(0.048)

Table 2: Estimation of Eigenfunctions and Measurement Error ($\delta = 0$)

N	δ	RASE_{v_1}	RASE_{v_2}	MSE	$\hat{\sigma}^2 = 0.01$
20	0.5	0.1682(0.0866)	0.2042(0.0624)	0.0102(0.0003)	0.0102(0.0003)
	0	0.1526(0.0684)	0.2042(0.0625)	0.0102(0.0003)	0.0102(0.0003)
40	0.5	0.1481(0.0855)	0.2122(0.0506)	0.0111(0.0003)	0.0111(0.0003)
	0	0.1394(0.0756)	0.2121(0.0506)	0.0111(0.0003)	0.0111(0.0003)

the estimation procedure 'incorporating correlation' selected similar optimal bandwidth $\hat{h}_\mu$ to the estimation procedure of 'working independent'. Hence, for the simplicity of our simulation study, we use the same bandwidth for the mean functions in the two estimation procedures. For the number of eigenfunctions, since both CV and pseudo-AIC did not work well in our simulation, we considered the rule-of-thumb criterion. In 500 simulations for both cases $\delta = 0$ and $\delta = 0.5$, the threshold of 85 percent explained variance selected the correct number of eigenfunctions for each component in more than 90% runs. For computational consideration we also assume that number of eigenfunctions are known in our simulation.

Table 1 displays the simulation results for both the cases of $\delta = 0.5$ and $\delta = 0$ over 500 simulations. The mean and standard deviation of RASE_μ, and the estimate of ρ_1 are recorded for both estimation procedures. The bandwidths are chosen as $(\hat{h}_\mu = 0.11, \hat{h}_{cov} = 0.10)$ when $N = 20$, and $(\hat{h}_\mu = 0.08, \hat{h}_{cov} = 0.08)$ when $N = 40$. For the $\delta = 0.5$ setting, the results show that the proposed procedures perform quite well for the selected bandwidths in the two

Table 3: Bootstrap standard error ($N = 20$, $\delta = 0.5$)

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mu_1(\cdot)$	SD	0.028	0.037	0.044	0.048	0.050	0.049	0.044	0.036	0.028
	SE	0.027	0.032	0.038	0.043	0.045	0.043	0.039	0.033	0.027
	Std	0.005	0.007	0.007	0.008	0.008	0.008	0.007	0.006	0.005
$\mu_2(\cdot)$	SD	0.032	0.025	0.025	0.031	0.020	0.033	0.026	0.026	0.031
	SE	0.033	0.025	0.024	0.031	0.019	0.032	0.024	0.025	0.033
	Std	0.005	0.004	0.004	0.005	0.004	0.005	0.005	0.004	0.004
$v_{11}(\cdot)$	SD	0.150	0.123	0.094	0.059	0.044	0.062	0.092	0.128	0.154
	SE	0.143	0.116	0.089	0.063	0.048	0.061	0.087	0.114	0.141
	Std	0.036	0.029	0.023	0.015	0.010	0.014	0.023	0.032	0.038
$v_{12}(\cdot)$	SD	0.096	0.115	0.138	0.170	0.174	0.149	0.143	0.115	0.089
	SE	0.111	0.119	0.140	0.157	0.164	0.158	0.143	0.122	0.112
	Std	0.020	0.019	0.025	0.031	0.033	0.030	0.026	0.020	0.021
$v_{21}(\cdot)$	SD	0.163	0.170	0.115	0.054	0.160	0.184	0.177	0.122	0.074
	SE	0.095	0.161	0.157	0.098	0.177	0.098	0.159	0.157	0.096
	Std	0.060	0.073	0.081	0.089	0.091	0.089	0.081	0.069	0.061
$v_{22}(\cdot)$	SD	0.198	0.108	0.173	0.158	0.123	0.189	0.120	0.169	0.182
	SE	0.229	0.181	0.181	0.234	0.203	0.237	0.180	0.183	0.226
	Std	0.057	0.034	0.046	0.043	0.044	0.054	0.036	0.046	0.053

estimation procedures. This suggests that when the components are well separated, the estimation procedure incorporating correlations does not provide significant improvements compared to the working independent procedure. For the $\delta = 0$ setting, the estimation procedure for working independent correlation performs quite poorly, and the estimate of proportion parameter ρ_1 has large bias. However, the estimation procedure incorporating correlations does give much better results: smaller RASE_μs for the mean functions, and more accurate estimates of ρ_1. The results agree with the explanations in the remark of Section 3.2.3 as expected. For the iterative estimation procedure, we further summarize the RASE of the eigenfunctions for each component, the MSE,

Table 4: Bootstrap standard error ($N = 40$, $\delta = 0$)

		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\mu_1(\cdot)$	SD	0.024	0.030	0.037	0.043	0.044	0.043	0.039	0.032	0.026
	SE	0.022	0.028	0.034	0.038	0.040	0.038	0.034	0.028	0.022
	Std	0.003	0.004	0.005	0.005	0.005	0.005	0.005	0.004	0.003
$\mu_2(\cdot)$	SD	0.037	0.026	0.027	0.037	0.017	0.037	0.025	0.027	0.036
	SE	0.035	0.027	0.027	0.035	0.020	0.035	0.026	0.027	0.035
	Std	0.004	0.004	0.005	0.006	0.005	0.005	0.005	0.004	0.005
$v_{11}(\cdot)$	SD	0.143	0.123	0.093	0.055	0.035	0.056	0.093	0.126	0.146
	SE	0.120	0.104	0.079	0.051	0.039	0.053	0.081	0.104	0.120
	Std	0.029	0.027	0.021	0.014	0.009	0.014	0.020	0.026	0.030
$v_{12}(\cdot)$	SD	0.072	0.110	0.135	0.156	0.164	0.154	0.130	0.104	0.074
	SE	0.087	0.104	0.124	0.139	0.144	0.139	0.123	0.102	0.084
	Std	0.013	0.015	0.022	0.025	0.026	0.026	0.022	0.016	0.013
$v_{21}(\cdot)$	SD	0.060	0.121	0.122	0.053	0.145	0.054	0.122	0.120	0.055
	SE	0.077	0.137	0.138	0.079	0.165	0.080	0.141	0.135	0.078
	Std	0.051	0.057	0.059	0.080	0.077	0.085	0.069	0.049	0.051
$v_{22}(\cdot)$	SD	0.153	0.105	0.105	0.150	0.063	0.153	0.105	0.110	0.147
	SE	0.185	0.145	0.145	0.192	0.137	0.196	0.146	0.147	0.184
	Std	0.075	0.068	0.068	0.090	0.117	0.110	0.084	0.063	0.068

and the estimate of σ^2 in Table 2. The results show that both the $\hat{\sigma}^2$ yielded by the iterative procedure and the MSE are good estimates of σ^2. In the heavy overlap setting, the proposed iterative procedure is able to provide good estimate of the eigenfunctions as well as the separated setting.

The accuracy of the standard error via bootstrap method can be assessed by Monte Carlo method. Table 3 and Table 4 summarize the performance of the standard errors of the mean functions and principal component functions at $t = 0.1, 0.2, \cdots, 0.9$. Denoted by SD the standard deviation of 200 estimates, which can be viewed as the true standard errors. The average and

Table 5: Comparisons for different error distributions

		Working independent		Incorporating correlation	
Distribution	δ	RASE_μ	$\rho_1 = 0.45$	RASE_μ	$\rho_1 = 0.45$
t(3)	0.5	0.062(0.014)	0.442(0.049)	0.066(0.021)	0.459(0.052)
	0	0.134(0.032)	0.301(0.050)	0.064(0.015)	0.485(0.055)
Laplace	0.5	0.060(0.013)	0.442(0.051)	0.059(0.012)	0.448(0.050)
	0	0.131(0.032)	0.305(0.049)	0.058(0.011)	0.466(0.053)
Exp(1)	0.5	0.060(0.014)	0.443(0.049)	0.058(0.012)	0.449(0.049)
	0	0.133(0.034)	0.303(0.051)	0.058(0.012)	0.466(0.052)

standard deviation of the 200 estimated standard errors via bootstrap, denoted by SE and Std, respectively, are recorded in rows. The result shows that the proposed standard error method works well for the mean functions and the eigenfunctions of the first component. However, it does not give very good result for the eigenfunctions of the second component. In simulation we use the same bandwidth h_{cov} in both covariances smoothing for simplicity of computation and bandwidth selection. The estimation may be improved by using different bandwidths in each component.

It is of interest to investigate whether the proposed model still works fine if the data do not follow Gaussian process. To this end, we consider three non-Gaussian distributions for the error term in model (2.1): (i) t-distribution with 3 degrees of freedom $0.1 \times t(3)$; (ii) Laplace distribution $0.1 \times Laplace(0, 1)$; (iii) centralized exp(1) distribution $0.1 \times (exp(1) - 1)$. In this simulation, we take the same setting as before except for the three error distributions. For the case $N = 20$, we report the mean and standard deviation of RASE_μ, and the estimate of ρ_1 over 500 simulations. The results summarized in Table 5 demonstrate that our estimation procedure is not very sensitive to the Gaussian assumption.

To investigate the performance of the proposed methodologies under large C, we conduct simulation studies by using $C = 20$ and 50. In the simulations, random observations are generated

Table 6: Simulation results for large C

		Working independent		Incorporating correlation	
C	δ^*	RASE_μ	$\|\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|\|$	RASE_μ	$\|\|\hat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|\|$
20	2	0.194(0.013)	0.030(0.005)	0.188(0.013)	0.030(0.005)
	4	0.186(0.013)	0.029(0.005)	0.186(0.013)	0.029(0.005)
50	2	0.415(0.031)	0.031(0.003)	0.402(0.026)	0.031(0.003)
	4	0.401(0.026)	0.031(0.003)	0.401(0.026)	0.031(0.003)

from a mixture of Gaussian processes with the following setting: $\rho_c = 1/C, \sigma^2 = 0.01$,

$$\mu_c(t) = \begin{cases} \sin(\pi t) + (c-1)\delta^*, & \text{if } c \text{ is odd,} \\ 1.5\sin(\pi t) + (c-1)\delta^* + 1, & \text{if } c \text{ is even.} \end{cases}$$

$$v_{1c}(t) = \begin{cases} \sqrt{2}\sin(\pi t), & \text{if } c \text{ is odd,} \\ \sqrt{2}\sin(4\pi t), & \text{if } c \text{ is even.} \end{cases}$$

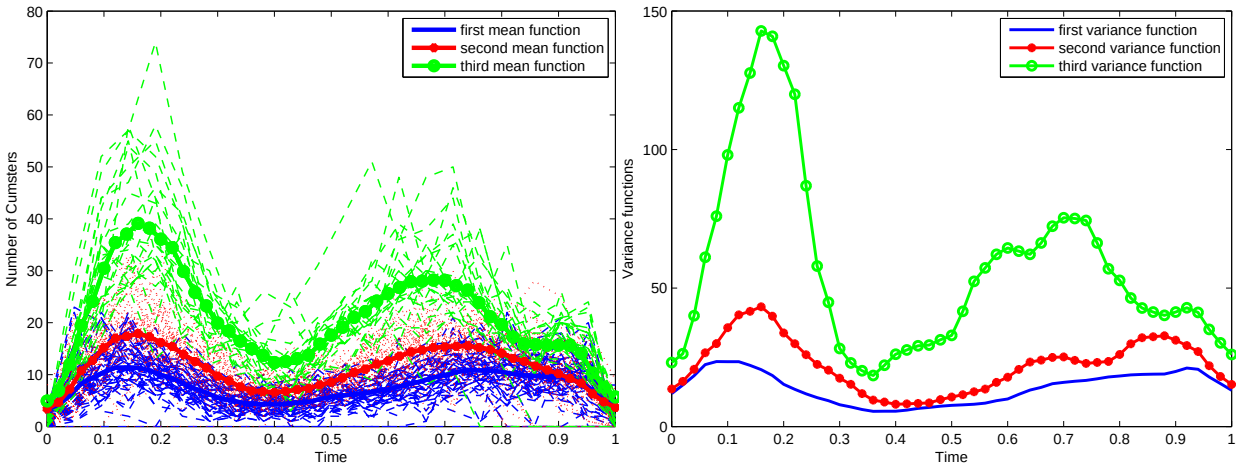
$$v_{2c}(t) = \begin{cases} \sqrt{2}\cos(\pi t), & \text{if } c \text{ is odd,} \\ \sqrt{2}\cos(4\pi t), & \text{if } c \text{ is even.} \end{cases}$$

The eigenvalues for all components are set as $\lambda_{1c} = 0.04$, $\lambda_{2c} = 0.01$, and $\lambda_{qc} = 0$, for $q > 2$. The principal component scores ξ_{iqc} are generated from $N(0, \lambda_{qc})$, $q = 1, 2$, and $c = 1, \cdots, C$.

For both cases $C = 20$ and 50, the simulated data with sample size $n = 1000$ are observed at grid points $\{k/N, k = 1, \cdots, N\}$ for both components, where $N = 20$. We consider two scenarios: $\delta^* = 4$ for well separated components, and $\delta^* = 2$ for heavily overlapping components. We ran 100 simulations for both scenarios, and the detailed results are given in Table 6. The results show that the proposed procedures still perform well when C is large.

4.2 Analysis of Supermarket Data

We use the proposed mixture of Gaussian processes and estimation procedure to analyze the supermarket dataset, which is depicted in Figure 1. We determine the number of component C using some partial sparse data. Since BIC often chooses simple models with finite sample, we consider the AIC for multivariate mixture of normals with one, two, three and four components.



(a) (b)

Figure 3: (a)Estimated mean functions and clustering results based on posteriors; (b)Estimated variance functions

We choose 4 sparse datasets, which are taken from the original data for every 4, 5, 6 time locations. The AIC scores achieve the minimum at $C = 3$ for all the sparse datasets; thus, it is reasonable to select a 3-component model for analysis.

We first analyze the data using the working independent correlation model (2.2) with three components. Without loss of information, we transform the time interval of the data to $[0, 1]$. The smoothing parameter chosen by CV selector is $h_\mu = 0.07$. The estimated proportions of the three components (from up to down) are 0.1632, 0.4311, and 0.4057. The estimated mean functions and a hard-clustering result are shown in Figure 3(a). The hard-clustering is obtained by assigning component identities according to the largest $r_{ic}, c = 1, \cdots, C$. From this result and the original data with actual calender dates, we found that the days in the upper class are mainly from the month of Chinese spring festivals. Most Saturdays and Sundays fall in the middle class, and the weekdays generally fall in the lower class. The estimated mean functions can be viewed as estimated average customer flows of the three classes. We observed that there are two peaks of customer flows for 3 components. The first peak occurs around 9:00 am in all components. The second peak occurs around 2:00 pm for the first component, and 3:00 pm for the second and third component. This pattern may indicate that people tend to buy earlier in the afternoon during the days of spring festival. We further plot the estimated variance functions of the three components in Figure 3(b). Combining Figure 3(a) and Figure 3(b), we observed that the variance functions

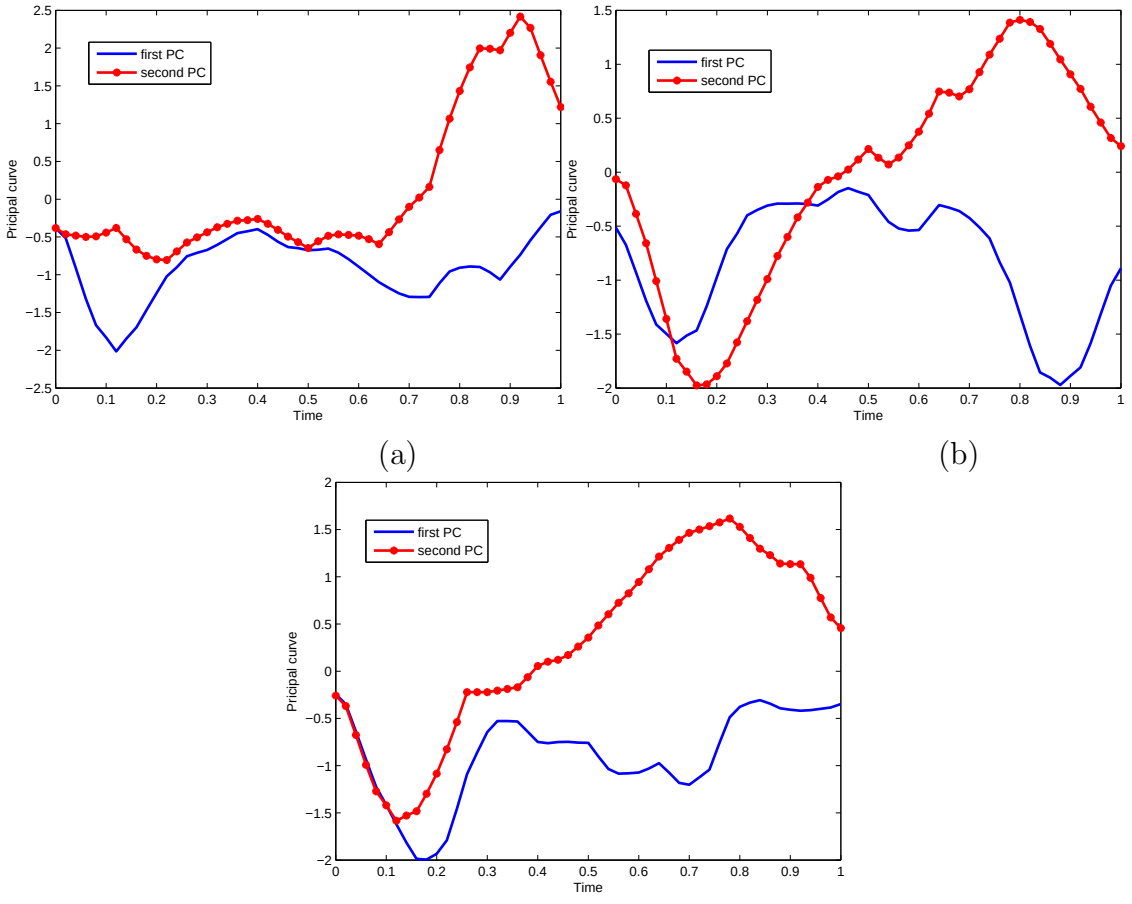


Figure 4: (a) First two eigenfunctions of the upper class; (b) First two eigenfunctions of the middle class; (c) First two eigenfunctions of the lower class.

followed a similar pattern with the mean functions in three components, in that a higher mean was associated with a higher variance.

The next step is to analyze the data by using functional principal component analysis. The selected bandwidth for the covariance function is $h_{cov} = 0.065$. Based on the estimated posterior, we estimate the covariance functions and obtain estimates of the eigenfunctions of all components. We plot the first two eigenfunctions of the three components in Figure 4. For the upper class, the first eigenfunction explains 51.70% of the total variation, and has a negative value along its time interval from 9:00 am to 5:30 pm. It means that a subject of this class (i.e., a day) with a positive (negative) functional principal component score on this direction tends to have smaller (larger) customer flows than the population average in a whole observed time interval. We also

observe that there are two negative peaks (corresponding to two lowest local minimums) in the first eigenfunction, which occurs around 9:00 am and 2:00 pm. It means that the variations of the customer flows are large in the two peaks, especially for the peak at 9:00 am. Note that these peaks are also observed in the first estimated variance function; therefore the results agree with each other as we expected. The second eigenfunction, which explains 22.80% of the total variation, has relatively small negative values in the morning and large positive values in the afternoon. This means that a subject with a positive functional principal component score on this direction tends to have smaller customer flow in the morning and a higher customer flow in the afternoon. The variation characterized by the second eigenfunction has a minor magnitude compared to the variation in the first eigenfunction, where the magnitude is determined by the eigenvalues. The third and fourth eigenfunction explains 7.58% and 4.28% of the total variation, and is of little interest. The first four principal components explain more than 85% percent of the total variation. Therefore, we think that using 4 eigenfunctions is enough for the analysis of the upper class. Similarly, we can analyze and interpret the eigenfunctions of the second component.

5 Discussion

Finite mixture models are particularly useful as a flexible modeling approach. In this paper, we proposed new estimation procedures for mixture of Gaussian processes. We imposed smoothed structures for both mean and covariance functions in each component, and showed that the mixture of Gaussian processes is identifiable under certain conditions. We further developed estimation procedures using kernel regression, EM algorithm, and functional principal component analysis. The proposed procedure overcomes several disadvantages of mixture of multivariate normals, such as "curse of dimensionality", and computational instability. It is easy to show that the computational complexities are $O(n \times N \times C \times n_{grid})$ and $O(n \times N^2 \times C \times n_{grid}^2 + C \times n_{grid}^3)$ for model (2.3) and model (2.1), respectively. The finite sample performance of the proposed method is examined by Monte Carlo simulation.

The selection of the number of components is a challenging problem. In this paper, we considered a computationally simple approach by fitting a multivariate normal mixtures to a partial data and demonstrated its effectiveness through supermaket data application. It requires further research to adaptively select the number of mixture components using some more complicated

methods. We may start with some likelihood-based approaches such as the information criterion method or penalized likelihood, however, a critical issue is to assess the model complexity, i.e., the effective number of parameters. In the nonparametric mixture of regression models, model complexity can be defined, e.g., Huang et al. (2013). However, in the proposed framework, there are still difficulties to obtain degree of freedom when we implement kernel regression and functional PCA for covariance estimation. Further researches on model complexity are needed. In addition to the primary interests of model estimation, testing in mixture models is also a very important issue. One may be interested in testing whether the mean functions are constant, or of a linear form. This issue can be further studied along the lines of nonparametric likelihood ratio test, e.g., Fan et al. (2001). It is interesting to study whether the Wilks Phenomenon still holds for the mixture of Gaussian processes.

In real application, data may not follow Gaussian process. We conducted some simulation to investigate whether the proposed model still works if the data do not follow Gaussian process. The results demonstrate that our method still works well when the error term in model (2.1) follows some other finite-moment distributions, such as t-distribution, Laplace distribution, and centralized exponential distribution. When there are additional functional covariate inputs, mixture of Gaussian process regression (Shi et al., 2005, 2007) can be used. It will be interesting to study how the proposed estimation methods in this article can be extended to the regression setting.

Appendix

Proof of Theorem 1

Suppose that $\{X(t), t \in T\}$ admits another representation such that given $\mathcal{D} = d, \{X(t), t \in T\}$ follows a Gaussian process with mean $\nu_d(t)$ and covariance function $\text{Cov}\{X(s), X(t)\} = H_d(s,t), d = 1, \ldots, D$. In addition, $P(\mathcal{D} = d) = \pi_d$. Therefore,

$$X(r) \sim \sum_{d=1}^{D} \pi_d N(\nu_d(r), H_d(r,r)) = \sum_{c=1}^{C} \rho_c N(\mu_c(r), G_c(r,r)).$$

Since the complement of $\mathbf{S}$ is not empty, there exists $r \in \mathbb{T}$ such that for any $1 \leq j \neq k \leq C$, $(\mu_j(r), G_j(r,r)) \neq (\mu_k(r), G_k(r,r))$. Based on the identifiability of finite mixture of normal distribution (see Titterington et al. (1985), p. 38, Example 3.1.4), $D = C$ and there exists a

permutation $\mathbf{w} = (w(1), \ldots, w(C))$ such that

$$\pi_{w(c)} = \rho_c, \nu_{w(c)}(r) = \mu_c(r), H_{w(c)}(r,r) = G_c(r,r), c = 1, \ldots, C. \tag{5.1}$$

Then for any pair (s,t) such that $r \neq s$, $r \neq t$, and $s \neq t$,

$$(X(r), X(s), X(t))^T \sim \sum_{c=1}^{C} \rho_c N_3 \left(\boldsymbol{\nu}_c(r,s,t), \mathbf{H}_c(r,s,t)\right) = \sum_{c=1}^{C} \pi_c N_3 \left(\boldsymbol{\mu}_c(r,s,t), \mathbf{G}_c(r,s,t)\right),$$

where

$$\boldsymbol{\nu}_c(r,s,t) = \begin{pmatrix} \nu_c(r) \\ \nu_c(s) \\ \nu_c(t) \end{pmatrix}, \quad \mathbf{H}_c(r,s,t) = \begin{pmatrix} H_c(r,r) & H_c(r,s) & H_c(r,t) \\ H_c(s,r) & H_c(s,s) & H_c(s,t) \\ H_c(t,r) & H_c(t,s) & H_c(t,t) \end{pmatrix},$$

$$\boldsymbol{\mu}_c(r,s,t) = \begin{pmatrix} \mu_c(r) \\ \mu_c(s) \\ \mu_c(t) \end{pmatrix}, \quad \mathbf{G}_c(r,s,t) = \begin{pmatrix} G_c(r,r) & G_c(r,s) & G_c(r,t) \\ G_c(s,r) & G_c(s,s) & G_c(s,t) \\ G_c(t,r) & G_c(t,s) & G_c(t,t) \end{pmatrix}.$$

Note that $(\mu_c(r), G_c(r,r))$s are different for different components. Based on Yakowitz and Spragins (1968), the above multivariate normal mixture model is identifiable. Therefore, there exists a permutation $\mathbf{w}_{s,t} = (w_{s,t}(1), \ldots, w_{s,t}(C))$ such that

$$\pi_{w_{s,t}(c)} = \rho_c, \boldsymbol{\nu}_{w_{s,t}(c)}(r,s,t) = \boldsymbol{\mu}_c(s,t), \mathbf{H}_{w_{s,t}(c)}(r,s,t) = \mathbf{G}_c(r,s,t),\ c = 1, \ldots, C.$$

Noting that $(\mu_c(r), G_c(r,r))$s are different for different components, based on (5.1),

$$w_{s,t}(c) = w(c), c = 1, \ldots, C, \text{ for any } (s,t),$$

where $w(\cdot)$ is defined in (5.1). Therefore, for any (s,t), such that $r \neq s$, $r \neq t$, and $s \neq t$, we have

$$\pi_{w(c)} = \rho_c, \nu_{w(c)}(t) = \mu_c(t), H_{w(c)}(s,t) = G_c(s,t), c = 1, \ldots, C. \tag{5.2}$$

In addition, since $\mu_c(\cdot)$ and $G_c(\cdot)$ are continuous functions, $\nu_{w(c)}(t) = \mu_c(t), H_{w(c)}(r,t) = G_c(r,t)$, and $H_{w(c)}(r,r) = G_c(r,r)$. Therefore, there exists a constant permutation $w = (w(1), \ldots, w(C))$, which is independent of (s,t), such that

$$\pi_{w(c)} = \rho_c, \nu_{w(c)}(r) = \mu_c(r), H_{w(c)}(r,s) = G_c(r,s), c = 1, \ldots, C. \tag{5.3}$$

This completes the proof of identifiability.

References

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics*, 29(1):153–193.

Hall, P., Müller, H., and Wang, J. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517.

Heard, N., Holmes, C., and Stephens, D. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes. *Journal of the American Statistical Association*, 101(473):18–29.

Huang, M., Li, R., and Wang, S. (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association*, 108:929 – 941.

James, G., Hastie, T., and Sugar, C. (2000). Principal component models for sparse functional data. *Biometrika*, 87(3):587–602.

James, G. and Sugar, C. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408.

Lin, X. and Carroll, R. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *Journal of the American Statistical Association*, pages 520–534.

Luan, Y. and 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. and Zhong, W. (2008). Penalized clustering of large-scale functional data with multiple covariates. *Journal of the American Statistical Association*, 103(482):625–636.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer, New York.

Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 233–243.

Shi, J. Q., Murray-Smith, R., Titterington, M., and Ab, J. Q. (2005). Hierarchical gaussian process mixtures for regression. *Statistics and Computing*, pages 31–41.

Shi, J. Q., Wang, B., Murray-Smith, R., and Titterington, D. M. (2007). Gaussian process functional regression modeling for batch data. *Biometrics*, 63(3):714–723.

Staniswalis, J. and Lee, J. (1998). Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, pages 1403–1418.

Titterington, D., Smith, A., Makov, U., et al. (1985). *Statistical analysis of finite mixture distributions*, volume 38. Wiley New York.

Yakowitz, S. and Spragins, J. (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214.

Yao, F., Müller, H., Clifford, A., Dueker, S., Follett, J., Lin, Y., Buchholz, B., and Vogel, J. (2003). Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. *Biometrics*, 59(3):676–685.

Yao, F., Müller, H., and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.