

# Functional mixture regression

FANG YAO\*

*Department of Statistics, University of Toronto, Toronto, Ontario M5S 3G3, Canada*  
fyao@utstat.toronto.edu

YUEJIAO FU

*Department of Mathematics and Statistics, York University, Toronto, Ontario M3J 1P3, Canada*

THOMAS C. M. LEE

*Department of Statistics, University of California, Davis, CA 95616, USA*

## SUMMARY

In functional linear models (FLMs), the relationship between the scalar response and the functional predictor process is often assumed to be identical for all subjects. Motivated by both practical and methodological considerations, we relax this assumption and propose a new class of functional regression models that allow the regression structure to vary for different groups of subjects. By projecting the predictor process onto its eigenspace, the new functional regression model is simplified to a framework that is similar to classical mixture regression models. This leads to the proposed approach named as functional mixture regression (FMR). The estimation of FMR can be readily carried out using existing software implemented for functional principal component analysis and mixture regression. The practical necessity and performance of FMR are illustrated through applications to a longevity analysis of female medflies and a human growth study. Theoretical investigations concerning the consistent estimation and prediction properties of FMR along with simulation experiments illustrating its empirical properties are presented in the supplementary material available at *Biostatistics* online. Corresponding results demonstrate that the proposed approach could potentially achieve substantial gains over traditional FLMs.

*Keywords:* Dimensional reduction; Eigenfunction; Functional data; Functional linear model; Functional principal components; Mixture regression; Smoothing.

## 1. INTRODUCTION

Recently, there has been an increased interest in regression models for functional data. In the simplest setting, the functional predictor and the scalar response are related by a linear operator. Given a scalar response  $Y$  on  $\tilde{R}$  and a smooth random predictor process  $X(\cdot)$  on a compact support  $\tilde{T}$  that is square integrable (i.e.,  $\int_{\tilde{T}} X^2(t)dt < \infty$ ), the classical functional linear model (FLM) relates  $Y$  and  $X$  by

$$E(Y|X) = \int_{\tilde{T}} \beta(t)X(t)dt, \quad (1.1)$$

\*To whom correspondence should be addressed.

where the regression parameter function  $\beta(\cdot)$  is also assumed to be smooth and square integrable. See Ramsay and Silverman (2005) for a comprehensive introduction. For further theoretical studies on (1.1), see Cardot and others (1999, 2003), Cai and Hall (2006) and Hall and Horowitz (2007).

Driven by the needs of generalizing the basic linear relationship in (1.1), several extensions of the above FLM have been proposed. This is similar to, for example, extending the classical linear regression models to generalized linear models. One of the early examples is generalized FLMs (Müller and Stadtmüller, 2005), and other extensions include varying-coefficient functional models (Fan and Zhang, 2000; Fan and others, 2003) and wavelet-based functional models (Morris and others, 2003).

In line with these extensions and motivated by the fact that, due to some unknown reasons or unobserved covariates, the subjects may belong to different mutually exclusive groups that possess different mechanisms to produce the response, we propose a new class of functional regression models. Our approach achieves this goal by allowing individuals from different groups to have distinct regression functions. To be specific, denote the unknown number of groups as  $K$ , and let  $\beta_k(t)$  be the regression function for the  $k$ th group,  $k = 1, \dots, K$ . Then, we propose the following model:

$$E(Y|X) = \int_{\mathcal{T}} \beta_k(t)X(t)dt \quad \text{if the subject belongs to the } k\text{th group.} \quad (1.2)$$

We shall first illustrate the utility of our proposal through an analysis of the biodemographic characteristics of female medflies (Müller and Zhang, 2005). This study concerns the dependence of longevity on the dynamics of the early fertility process and we shall show that our proposal sheds new light on various important scientific issues. The second example, derived from the Berkeley growth study, considers the regression of heights at maturity age on the childhood growth patterns. We shall illustrate how distinct regression relations emerge and reveal the underlying gender groups, even when we were completely blinded from gender information throughout the analysis. Our proposed method can also be potentially useful in various medical applications, such as when  $X(t)$  is a longitudinal biomarker and  $Y$  is a disease indicating variable; for example, glomerular filtration rate in kidney diseases and postload glucose in type 2 diabetes.

Extending the classical FLM (1.1)–(1.2) is parallel to extending the classical linear regression to mixture regression (DeSarbo and Cron, 1988), thus termed as functional mixture regression (FMR). We emphasize that a main goal of FMR is to specify an appropriate functional model that is capable of identifying potentially different regression structures. This general idea can readily be adopted to various applications in which functional regression techniques are needed. We also remark that FMR is conceptually different from existing approaches for curve-based clustering (Gaffney and Smyth, 2003; James and Sugar, 2003; Luan and Li, 2003, among others). These latter methods focus on clustering the trajectories themselves, while FMR focuses on detecting the possible existence of different regression relations.

The rest of this paper is organized as follows. In Section 2, we provide a complete description of FMR and demonstrate that its estimation can be achieved using existing software implemented for functional principal component analysis (FPCA) and mixture regression. Applications of the proposed method to the above-mentioned real examples are presented in Section 3. Concluding remarks are offered in Section 4, while simulations illustrating the empirical performance and theoretical investigation on consistent estimation and prediction are deferred to supplementary material available at *Biostatistics* online for conciseness.

## 2. FUNCTIONAL MIXTURE REGRESSION

We begin with the classical FLM (1.1) and review a key methodology for dimension reduction and regularization of functional data, namely, FPCA. For introductory material on FPCA, see Rice and Silverman (1991), James and others (2001), Ramsay and Silverman (2005), Yao and others (2005), among others.

### 2.1 FLM and eigenbasis representation

We begin with the classical FLM (1.1), where the regression function  $\beta(t)$  is the same for all subjects under consideration. It is known that the functional linear operator in (1.1) is compact and not directly invertible, and thus regularization is needed and can be achieved through a truncated basis representation. In this article, we shall adopt the eigenbasis representation for reasons to be given below.

The process  $X$  with finite covariance possesses a sequence of orthonormal eigenfunctions  $\{\phi_m\}_{m=1,2,\dots}$ , which form a complete basis of the functional space, with associated nonnegative and nondecreasing eigenvalues  $\{\lambda_m\}_{m=1,2,\dots}$ . By the well-known Karhunen–Loève expansion, the predictor process  $X$  admits

$$X(t) = \mu(t) + \sum_{m=1}^{\infty} \zeta_m \phi_m(t), \quad \text{where } \zeta_m = \int_{\mathcal{T}} \{X(t) - \mu(t)\} \phi_m(t) dt. \quad (2.1)$$

The random variables  $\zeta_m$ s are the functional principal component (FPC) scores of  $X$ , which are uncorrelated and satisfy  $E(\zeta_m) = 0$  and  $\text{var}(\zeta_m) = \lambda_m$ ,  $\sum_m \lambda_m < \infty$ . The trajectory  $X_i$  is an i.i.d. realization of  $X$  for the  $i$ th subject related to the response  $Y_i$ ,  $i = 1, \dots, n$ . Here, we highlight the following advantages of (2.1) for regression regularization that will also be carried over to FMR. First, the eigenfunctions are orthonormal in the  $L^2$  space and the FPC scores are uncorrelated random variables, which provide both analytical and computational convenience. Second, the eigenbasis are determined by the data and will efficiently capture the dominant modes of variation. The eigenvalues often decrease rapidly and thus the infinite-dimensional predictor process can be well approximated by a small number of FPCs. This suggests a simple way to achieve regularization by truncating eigenbasis based on the total variation explained up to a threshold.

Recall that the regression parameter function  $\beta$  is square integrable and  $\{\phi_m\}_{m=1,2,\dots}$  form a complete orthonormal basis, we have  $\beta(t) = \sum_{m=1}^{\infty} b_m \phi_m(t)$ , and hence model (1.1) can be expressed equivalently as

$$\begin{aligned} E(Y|X) &= \int_{\tilde{\mathcal{T}}} \beta(t) \mu(t) dt + \int_{\tilde{\mathcal{T}}} \left\{ \sum_{m=1}^{\infty} b_m \phi_m(t) \right\} \left\{ \sum_{m=1}^{\infty} \zeta_m \phi_m(t) \right\} dt \\ &= b_0 + \sum_{m=1}^{\infty} b_m \zeta_m, \end{aligned} \quad (2.2)$$

where the intercept is  $b_0$  and the coefficients are given by  $b_m = \int_{\tilde{\mathcal{T}}} \beta(t) \phi_m(t) dt$ . One can see that the orthonormality of the complete eigenbasis plays a critical role in transforming the functional regression structure into a linear combination of the uncorrelated FPC scores that serve as predictor variables in (2.2).

### 2.2 Model specification of functional mixture regression

This section provides a complete mathematical formulation of FMR. Recall that FMR allows the predictor trajectories  $X_i$  to partition into  $K$  mutually exclusive groups, with each group having its own regression function  $\beta_k(t)$  for producing the response  $Y_i$ . This idea was previously expressed by model (1.2) and it is useful to rewrite it in a similar manner as (2.2). In this paper,  $K$  is unknown and will be chosen by some statistical model selection criterion addressed later. Write  $\tilde{\mathcal{S}} = \{1, \dots, n\}$  and define the index set

$$\tilde{\mathcal{C}}_k = \{i \in \tilde{\mathcal{S}}: \text{the } i\text{th subject belongs to the } k\text{th group}\}, \quad k = 1, \dots, K.$$

Thus,  $\bigcup_{k=1}^K \tilde{\mathcal{C}}_k = \tilde{\mathcal{S}}$  and  $\tilde{\mathcal{C}}(k_1) \cap \tilde{\mathcal{C}}(k_2) = \emptyset$  for  $1 \leq k_1 \neq k_2 \leq K$ . Write  $b_{k0} = \int_{\mathcal{T}} \beta_k(t) \mu(t) dt$  and  $b_{km} = \int_{\mathcal{T}} \beta_k(t) \phi_m(t) dt$ . By analogy to (2.2), we have

$$\beta_k(t) = \sum_{m=1}^{\infty} b_{km} \phi_m(t), \quad t \in \mathcal{T}, \quad k = 1, \dots, K,$$

and the FMR model (1.2) can be written as

$$E(Y_i|X_i, i \in \tilde{C}_k) = b_{k0} + \sum_{m=1}^{\infty} b_{km}\zeta_{im}. \tag{2.3}$$

We note that in (2.3) the FPC scores  $\zeta_{im}$  serve as the predictor variables and that the infinite-dimensional feature is inherent to the functional data. The estimation of model (2.3) thus requires regularization for the predictor process  $X$ , and we achieve this by truncating the infinite sum to a finite sum of  $M$  terms. Since often the eigenvalues of  $X$  decrease to zero rapidly, it is reasonable to assume that an appropriate  $M$  can always be chosen that leads to a flexible yet parsimonious model. A simple and practical strategy is to examine the total variation explained up to certain threshold  $\tau$ ,  $M = \min \{ \ell: \sum_{m=1}^{\ell} \lambda_m / \sum_{m=1}^{\infty} \lambda_m \geq \tau \}$ . With this truncation, model (2.3) is refined,

$$E(Y_i|X_i, M, i \in \tilde{C}_k) = b_{k0} + \sum_{m=1}^M b_{km}\zeta_{im}, \tag{2.4}$$

as the underlying model and the regression parameter functions that we aim for are

$$\beta_{k,M}(t) = \sum_{m=1}^M b_{km}\phi_m(t) \quad \text{for all } t \in \mathcal{T} \text{ and } k = 1, \dots, K. \tag{2.5}$$

From (2.4), the FMR model is now reduced to a form similar to classical mixture of linear regression models.

To complete the mathematical description of the FMR model (2.4) from which the statistical inference is based on, let  $\pi_k$  be the probability that a randomly selected subject is from the  $k$ th group and define  $\sigma_{ky}^2 = \text{var}(Y_i|X_i)$  if  $i \in \tilde{C}_k$ . Write  $\mathbf{y} = (y_1, \dots, y_n)^T$ ,  $\mathbf{b}_0 = (b_{10}, \dots, b_{K0})^T$ ,  $\mathbf{b}_k = (b_{k1}, \dots, b_{kM})^T$ ,  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{K-1})^T$ , and  $\boldsymbol{\sigma}_y^2 = (\sigma_{1y}^2, \dots, \sigma_{Ky}^2)^T$ . Denote  $\boldsymbol{\psi} = (\mathbf{b}_0^T, \mathbf{b}_1^T, \dots, \mathbf{b}_K^T, \boldsymbol{\pi}^T, \boldsymbol{\sigma}_y^T)^T$ , the parameter space  $\Theta$  is then given by

$$\Theta \equiv \left\{ \boldsymbol{\psi}: \pi_k > 0, \sum_{k=1}^K \pi_k = 1, \sigma_{ky}^2 > 0 \text{ for all } k = 1, \dots, K \right\}, \tag{2.6}$$

which is an open subset of  $\tilde{R}^{(M+3)K-1}$ . Further, write  $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iM})^T$  for  $i = 1, \dots, n$ ; these vectors  $\boldsymbol{\xi}_i$ s of the FPC scores are i.i.d. realizations of a random vector  $\boldsymbol{\xi}$  whose density is  $f(\boldsymbol{\xi}|\Lambda)$  with  $E(\boldsymbol{\xi}) = \mathbf{0}$  and  $\text{cov}(\boldsymbol{\xi}, \boldsymbol{\xi}) = \text{diag}\{\lambda_1, \dots, \lambda_M\} \equiv \Lambda$ . The conditional density of  $Y$  given  $\boldsymbol{\xi}$  is

$$f(y|\boldsymbol{\xi}, \boldsymbol{\psi}) = \sum_{k=1}^K \pi_k f(y|\boldsymbol{\xi}, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2), \tag{2.7}$$

where  $f(y|\boldsymbol{\xi}, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  is the conditional density for the  $k$ th component. In general, the component density can be derived from a location-scale family (Hennig, 2000) or an exponential family (Wedel and DeSarbo, 1995) that generates identifiable mixtures. This includes most commonly adopted distributions, such as normal, gamma, exponential, Poisson, binomial, and multinomial. To emphasize the main idea of coupling functional data, we focus on the mixture of linear regressions with normal errors. It is conceptually straightforward to extend our proposal to the mixture of generalized linear models with estimation procedure modified accordingly. Note that the above formulation holds for model (2.4) and the dependence on  $M$  is suppressed for.

Since the  $\boldsymbol{\xi}_i$ s are random, (2.4) becomes a random design regression model. As to be described in Section 2.3, the matrix  $\Lambda$  will be estimated in a way that is functionally independent of  $\boldsymbol{\psi}$ , therefore, the inference of (2.4) is completely based on the conditional density  $f(y|\boldsymbol{\xi}, \boldsymbol{\psi})$ . As  $\lambda_1 \geq \dots \geq \lambda_M > 0$  for

any  $M$ , it is easy to see that  $f(\xi|\Lambda)$  does not have all its mass in up to  $K$  of  $(M - 1)$ -dimensional linear subspaces. This implies that the FMR model (2.4) is identifiable in the following sense (Hennig, 2000): for any 2 parameters  $\psi$  and  $\psi^*$  with a given predictor variable  $\xi_i$ , if

$$\sum_{k=1}^K \pi_k f(y|\xi_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \sum_{k=1}^{K^*} \pi_k^* f(y|\xi_i, b_{k0}^*, \mathbf{b}_k^*, \sigma_{ky}^{*2}) \quad (2.8)$$

is true for all  $i$  and all possible values of  $y$ , then  $K = K^*$  and  $\psi = \psi^*$  up to a permutation. It is noteworthy that an attractive property of (2.4) is that its predictor variables  $\xi_{im}$ s are uncorrelated, hence devoid of collinearity.

It is important to note that the condition that  $X_i$  are i.i.d. processes does not exclude the possibility of  $\xi_i$  itself following a mixture distribution because the Karhunen–Loève expansion (2.1) that FPC analysis is based on only requires the existence of covariance function. The inference on the regression structure is based on the conditional density (2.7) and does not depend on the distribution of  $\xi_i$ . For instance, suppose that  $\xi_i$ s follow a mixture density. The regression structure does not necessarily vary across the groups of subjects partitioned by the distribution of  $\xi_i$ . In other words, the assignments of group membership for an individual in FMR is determined merely by the relationship between  $X_i$  and  $Y_i$ . Another noteworthy remark is that, for the reason of detecting different regression structures, the mixing proportion  $\pi_k$  is assumed independent of the predictor  $X_i$ , which is similar to classical mixture regression models. This is distinct from the class of hierarchical mixture of experts arising from the neural network literature (Jiang and Tanner, 1999), where it is common to assume that the  $\pi_k$ s depend on covariates to allow for flexible approximation of the overall mean response function.

### 2.3 Model estimation and implementation

This subsection discusses approaches for estimating the unknowns in (2.4) and (2.5), which can be naturally done in 2 stages. Briefly, in the first stage, we perform FPCA to obtain estimates for  $\phi_m$  and  $\xi_{im}$ , while in the second stage, these estimates are plugged into (2.4) and (2.5) for the estimation of the remaining parameters.

In practice, the observed data are noisy measurements  $U_{ij}$  taken at  $t_{ij}$ ,

$$U_{ij} = X_i(t_{ij}) + \varepsilon_{ij} = \mu(t_{ij}) + \sum_{m=1}^{\infty} \xi_{im} \phi_m(t_{ij}) + \varepsilon_{ij}, \quad t_{ij} \in \mathcal{T}, \quad (2.9)$$

for  $i = 1, \dots, n$  and  $j = 1, \dots, n_i$ . The measurement errors  $\varepsilon_{ij}$  are assumed independent of  $\xi_{im}$  with mean zero and a constant variance  $E\varepsilon_{ij}^2 = \sigma_x^2$ , while a nonconstant variance function could also be assumed to account for heteroscedasity (Yao and Lee, 2006). We first apply the principal analysis by conditional estimation (PACE) procedure of Yao and others (2005) to these noisy measurements to carry out FPCA. When this is done, the following estimates of model components are obtained:  $\hat{\mu}$ ,  $\hat{G}$ ,  $\hat{\phi}_m$ ,  $\hat{\lambda}_m$ ,  $\hat{\xi}_{im}$ ,  $m = 1, \dots, M$ . Here,  $M$  is the number of FPCs that can be chosen by pseudo-Akaike Information Criterion or other related selectors or simply as the minimum number of FPCs that explain a sufficiently large proportion of the total variation for the predictor process. We adopted the latter approach and found that the 90% threshold works excellently for our numerical examples. In general, one may need to navigate several choices of the threshold values to determine the model that provides an adequate fit with parsimonious structure.

For conciseness, we refer to Yao and others (2005) for a complete description of the FPCA technique used in this paper. Here, we only present the integral and PACE estimates of the FPC scores  $\xi_{im}$  with the

notation introduced in Section 2.1. The integral estimate is given by

$$\hat{\xi}_{im}^I = \sum_{j=2}^{n_i} (U_{ij} - \hat{\mu}(t_{ij})) \hat{\phi}_m(t_{ij})(t_{ij} - t_{i,j-1}), \quad (2.10)$$

which is motivated by the definition of the FPC scores as inner products; that is,  $\xi_{im} = \int_T \{X_i(t) - \mu(t)\} \phi_m(t) dt$ . For the PACE estimates, write  $\boldsymbol{\mu}_i = (\mu(t_{i1}), \dots, \mu(t_{in_i}))^T$ ,  $\mathbf{U}_i = (U_{i1}, \dots, U_{in_i})^T$ , and  $\boldsymbol{\phi}_{im} = (\phi_m(t_{i1}), \dots, \phi_m(t_{in_i}))^T$ , and let the  $(j, l)$ -th entry of the  $n_i \times n_i$  matrix  $\Sigma_{U_i}$  be  $(\Sigma_{U_i})_{j,l} = G(t_{ij}, t_{il}) + \sigma_x^2 \delta_{jl}$  with  $\delta_{jl} = 1$  if  $j = l$  and 0 if  $j \neq l$ . Substituting estimates for  $\boldsymbol{\mu}_i$ ,  $\lambda_m$ ,  $\boldsymbol{\phi}_{im}$ , and  $\Sigma_{U_i}$ , we have the PACE estimates

$$\hat{\xi}_{im}^P = \hat{\lambda}_m \hat{\boldsymbol{\phi}}_{im}^T \hat{\Sigma}_{U_i}^{-1} (\mathbf{U}_i - \hat{\boldsymbol{\mu}}_i). \quad (2.11)$$

It is widely known that when the design points  $t_{ij}$  are dense, the traditional integral estimates of the FPC scores  $\xi_{im}$ , denoted by  $\hat{\xi}_{im}^I$  and is given by (2.10) below, are usually satisfactory. By contrast, the PACE estimates  $\hat{\xi}_{im}^P$  as in (2.11) is more suitable when the design points are moderate or sparse. Corresponding software is available at <http://www.utstat.toronto.edu/fyao>.

Once the crucial estimates  $\hat{\xi}_{im}$ s of the FPC scores are obtained by either (2.10) or (2.11), the regression coefficients  $b_{km}$ s in (2.4) can be estimated in a relatively straightforward manner: with  $\hat{\xi}_{im}$  as the predictor variables, then  $b_{km}$ s can be estimated by standard mixture regression estimation method (e.g., expectation maximization-based method). The fitted FMR model and regression functions are then given by, for  $k = 1, \dots, K$ ,

$$\widehat{E}(Y_i | X_i, M) = \hat{b}_{k0} + \sum_{m=1}^M \hat{b}_{km} \hat{\xi}_{im}, \quad \hat{\beta}_{k,M}(t) = \sum_{m=1}^M \hat{b}_{km} \hat{\phi}_m(t) \quad \text{if } i \in \tilde{C}_k. \quad (2.12)$$

Notice that the estimated FPC scores  $\hat{\xi}_{im}$  are unique up to a sign change related to the direction of the estimated eigenfunctions  $\hat{\phi}_m(t)$ . This property is also carried over to the estimated regression coefficients  $\hat{b}_{km}$ . For the choice of  $K$ , one could apply any well-studied model selection criterion, and we adopt the Bayesian information criterion (BIC) that has provided good results in a variety of applications of model-based clustering (e.g., see Fraley and Raftery, 2002, and references therein).

For conducting inference procedures on the regression function(s), we could exploit nonparametric bootstrap methods with a suitable label-switching strategy for mixture regression to avoid nonidentifiability of component labels. More specifically, we first resample all the individuals with replacement to obtain a bootstrap sample,  $\{(U_{i1}^b, \dots, U_{in_i}^b, y_i^b) : i = 1, \dots, n, b = 1, \dots, B\}$ , and perform the FPCA step, where  $B$  is the number of bootstrap replicates. Then the estimated FPC scores  $\{(\hat{\xi}_{i1}^b, \dots, \hat{\xi}_{iM^b}^b) : i = 1, \dots, n\}$  are fed into a mixture regression model with  $K$  components, where  $M^b$  is chosen in FPCA using the same criterion as for the original sample, and  $K$  is the number of mixture components selected for the original sample. To correctly label the latent groups, we examine the distances from the estimated regression functions of the bootstrap sample,  $\hat{\beta}_\ell^b(t) = \sum_{m=1}^{M^b} \hat{b}_{\ell m}^b \hat{\phi}_\ell^b(t)$ , to those obtained from the original sample, that is,  $\hat{\beta}_k^b = \operatorname{argmin}_{\{\hat{\beta}_\ell^b : \ell=1, \dots, K\}} \int_T [\hat{\beta}_\ell^b(t) - \hat{\beta}_k(t)]^2 dt$  for each  $k = 1, \dots, K$ . We note that this bootstrap procedure also provides evidence for model identifiability.

We have derived theoretical results in terms of consistency of model estimation and prediction for FMR. In establishing such results, a first technical difficulty encountered is the fact that the estimates of the regression functions in the FMR model (2.4) are based on the estimated FPC scores  $\hat{\xi}_{im}$  not on the "true"  $\xi_{im}$ . Thus, existing theories of mixture regression models are no longer applicable. Another major challenge is due to the lack of analytic expressions for  $\hat{b}_{km}$ . Therefore, customary theoretical arguments previously used in FLMs cannot be applied. Due to space limitation, these technical contents such as the relevant theorems, assumptions, auxiliary lemmas, and proofs to the supplementary material available at *Biostatistics* online.

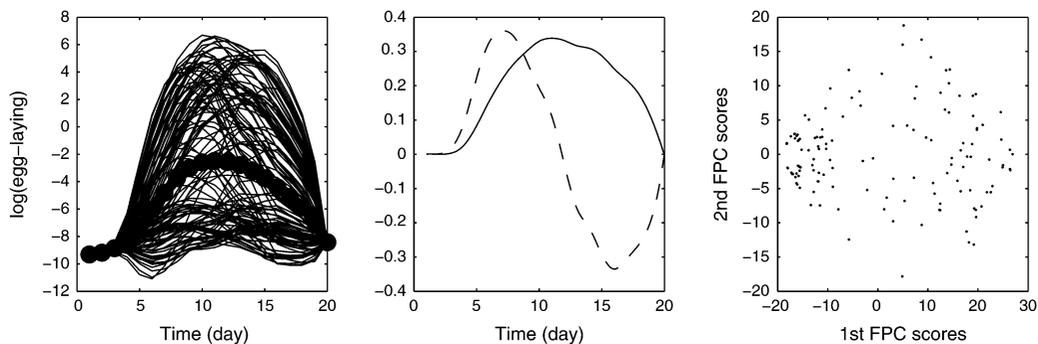
## 3. APPLICATIONS

## 3.1 Longevity and early fertility of mediterranean flies

To illustrate the need of the proposed approach, we analyze the egg-laying data from a fertility study conducted for 1000 female medflies as described in [Carey and others \(1998\)](#). Our goal is to determine the dependence of longevity of the medflies on their early fertility process. One of the basic questions of evolutionary theory is to what extent lifespan is driven by enabling increased reproduction. Diverting resources used for maintenance and repair into reproductive activity may shorten lifespan ([Partridge and Harvey, 1985](#); [Westendorp and Kirkwood, 1999](#)). The selected sample of 139 medflies includes those that were fertile during an early life period defined by the first 20 days and also survived beyond. The trajectories corresponding to the number of daily eggs during this early life period constitute the functional predictors, while remaining lifetime serves as the response that is an important proxy for longevity and quantifying the evolutionary fitness of individual flies. As a preprocessing step to achieve homogeneity, a log-transform of egg counts was applied.

These predictor trajectories (obtained by applying the PACE algorithm in FPCA step) are shown in the left panel of [Figure 1](#). Most egg-laying trajectories display a rise toward a time of peak fertility followed by a decline. There is substantial variation in the steepness of the rise to the various maximal level of egg-laying and also in the timing of the peak and the rate of decline. The smooth estimate of the mean fertility function is also displayed, while the estimates of the first 2 eigenfunctions are shown in the right panel, explaining 76.8% and 14.5% of the total variation of the trajectories, respectively. These eigenfunctions reflect the modes of variation ([Castro and others, 1986](#)) and the dynamics of predictor processes. Two components were chosen, and they account for more than 90% of the variation in the data, that is,  $\tau = 0.9$ .

It is of interest to identify shape changes in early life reproductive trajectories that tend to influence evolutionary longevity. To conduct an adequate analysis, we would inspect whether the regression relationship varies due to some unknown mechanism. It is noticed that there is no obvious grouping effect in the predictor trajectories observed. This can be seen from the perspective of the estimated FPC scores (right panel of [Figure 1](#)) that are often viewed as subject-specific summaries. However, when the remaining lifetimes are graphed versus the FPC scores in the right panel of [Figure 2](#), the lifespan seems driven by the early fertility differently with considerably longer lifetimes for some flies whose predictor patterns (in the left panel) might be similar. To verify this conjecture, we applied the FMR approach and unsurprisingly 2 mutually exclusive groups with different regression structures were suggested by BIC ( $K = 2$ ),



[Fig. 1](#). Left: smoothed egg-laying trajectories (functional predictor) for the 139 included flies with the smooth estimate of the mean function (thick solid curve). Middle: the first (solid) and second (dashed) estimated eigenfunctions explaining 76.8% and 14.5% of the total variation, respectively. Right: the estimated FPC scores obtained by the integral method (2.10) for the 139 flies.

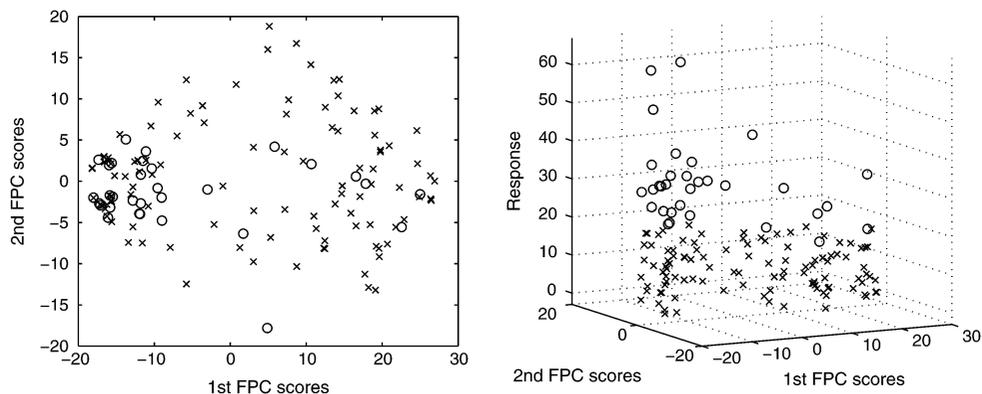


Fig. 2. Left: estimated FPC scores for the 139 flies with different markers (circles and crosses) used for representing the 2 mutually exclusive groups with different regressions as detected by FMR. Right: responses  $Y_i$  (remaining lifetime in day) against estimated FPC scores for the 139 flies with corresponding group assignments.

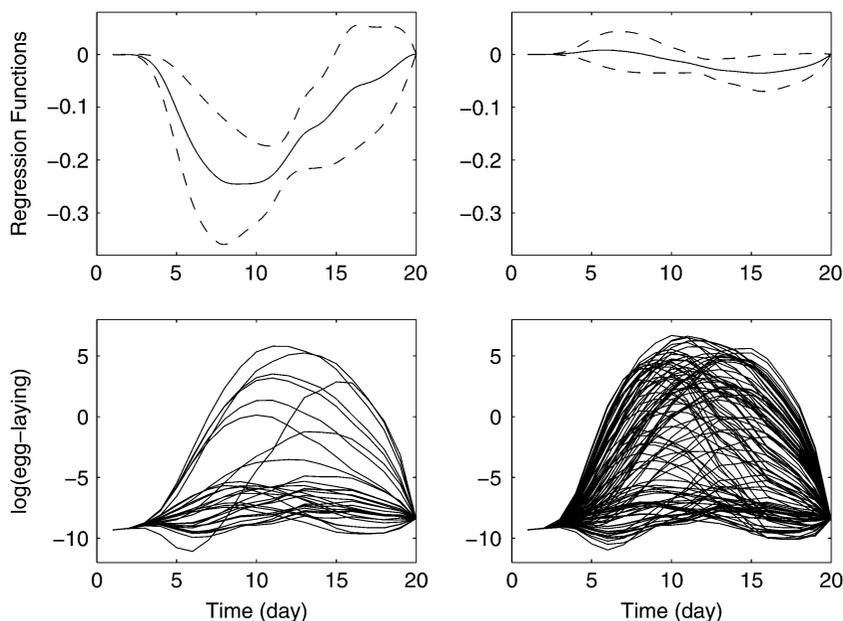


Fig. 3. Top panels: estimated regression functions  $\hat{\beta}_1$  (solid in top left) and  $\hat{\beta}_2$  (solid in top right) of the 2 groups detected by FMR along with 95% bootstrap confidence bands (dashed). Bottom panels: predictor trajectories of the flies (indicated by circles in the right panel of Figure 2) that correspond to  $\hat{\beta}_1$  (bottom left) and of the flies (indicated by crosses) that correspond to  $\hat{\beta}_2$  (bottom right).

where different markers (circles and crosses) were used for enhanced visualization of such phenomenon in the right panel of Figure 2.

The estimates of regression functions  $\beta_1$  (solid) and  $\beta_2$  (solid) serving as weighting functions shown in the top panels of Figure 3 indicate how the lifespan is influenced by the early fertility process, depending on which group a fly belongs to. We applied the nonparametric bootstrap procedure as described in Section 2.3 and constructed the 95% bootstrap confidence bands by taking 2.5th and 97.5th quantiles of

1000 replicates shown in the same panels. This provides a measure of accuracy of our point estimation and evidence for model identifiability as well. For illustration, we plotted again the predictor trajectories of 2 groups separately in bottom panels of Figure 3, where the flies in the middle panel ( $n_1 = 32$ ) corresponds to  $\beta_1$  that have longer lifetimes (indicated by circles in the right panel of Figure 2) and those in the right panel ( $n_2 = 107$ ) to  $\beta_2$  those live shorter (indicated by crosses). Overall higher level of fertility seems to shorten lifespan. More specifically, if a fly belongs to the group according to  $\beta_1$ , a slow rise to a lower peak of egg production helps to prolong the lifespan. By contrast, the flies in the other group with larger reproductivity around day 15 often have shorter lifetimes. These findings shed some new insight by distinguishing distinct underlying mechanisms relating longevity and early fertility. This may help experimenters look into evolutionary interpretation and implication of these mechanisms for different medflies. We conclude this example by a comparison with a FLM, where the leave-one-subject-out cross-validated relative prediction errors  $\text{CVRPE} = \sum_{i=1}^n (Y_i - \hat{Y}_i^{(-i)})^2 / \sum_{i=1}^n Y_i^2$  were obtained for FMR as 0.163 and for FLM as 0.372, indicating a substantial gain of 56% in prediction ability.

### 3.2 Berkeley growth study

Studies of human growth dynamics are an important topic in biological and medical applications that have profound impact for many years. This example concerns the Berkeley growth data originally published in [Tuddenham and Snyder \(1954\)](#) and analyzed by [Ramsay and others \(1995\)](#) in terms of height acceleration to reveal the dynamics of human growth. Similar data, for example, the Zurich growth data, were also studied from this perspective using various smoothing approaches ([Gasser and others, 1984](#), among others). It is known that the growth patterns of boys and girls during their pubertal spurts differ significantly in terms of magnitude and timing. Mainly for demonstration purpose, in this example, we study the human growth from a different perspective by examining the dependence of the height at maturity age 18 (scalar response) on the dynamic pattern till age 9 (predictor process) before pubertal spurts (see Figure 4).

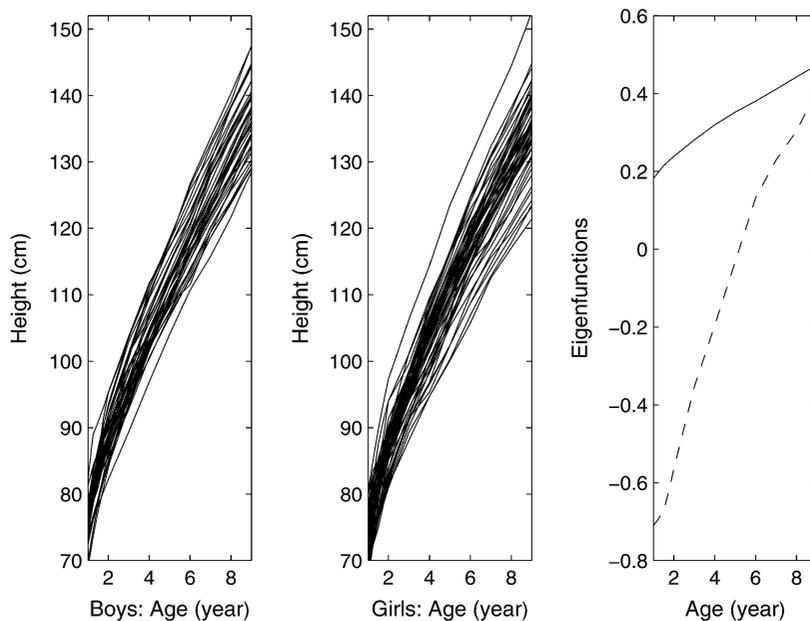


Fig. 4. Left and middle: height trajectories from age 1 to 9 for 39 boys (left) and 54 girls (middle). Right: smooth estimates of the first (solid) and second (dashed) eigenfunctions, accounting for 88.7% and 9.8% of total variation.

The data analyzed consist of height records for 39 boys and 54 girls, where the measurements were taken quarterly from ages 1 to 2, annually from 2 to 8, and semiannually from 8 till 18. It is worth mentioning that, for illustration, we shall blind ourselves from the gender information throughout the analysis, that is, the gender is an unknown or hidden factor that we expect the proposed approach is capable to detect.

We first carried out FPCA for the predictor process by pooling 93 trajectories together. Shown in the right panel of Figure 4 are the smooth estimates of the first 2 eigenfunctions account for 88.7% and 9.8% of the total variation, respectively, where the first eigenfunction is in the direction of the overall trend and the second shows a contrast between early and late times. The estimated FPC scores displayed in the top left panel of Figure 5 do not show a strong separation between boys and girls. However, when we examined the plot of the response  $Y_i$  against the estimated FPC scores in the top right panel, the separation between boys and girls becomes more apparent (circles for boys and crosses for girls). This phenomenon seems to suggest different regression relations for each gender.

The FMR approach indeed worked beautifully and led to 2 distinct groups based on BIC, that is,  $K = 2$ . Moreover, the partition based on FMR results corresponds to the gender group as expected in which only 1 boy and 2 girls were misclassified when we inspected the cross-validated classification. The regression functions for boys and girls are displayed in the bottom panels of Figure 5 along with the 95% bootstrap confidence bands. Recall that the first eigenfunction is an overall shift. The common

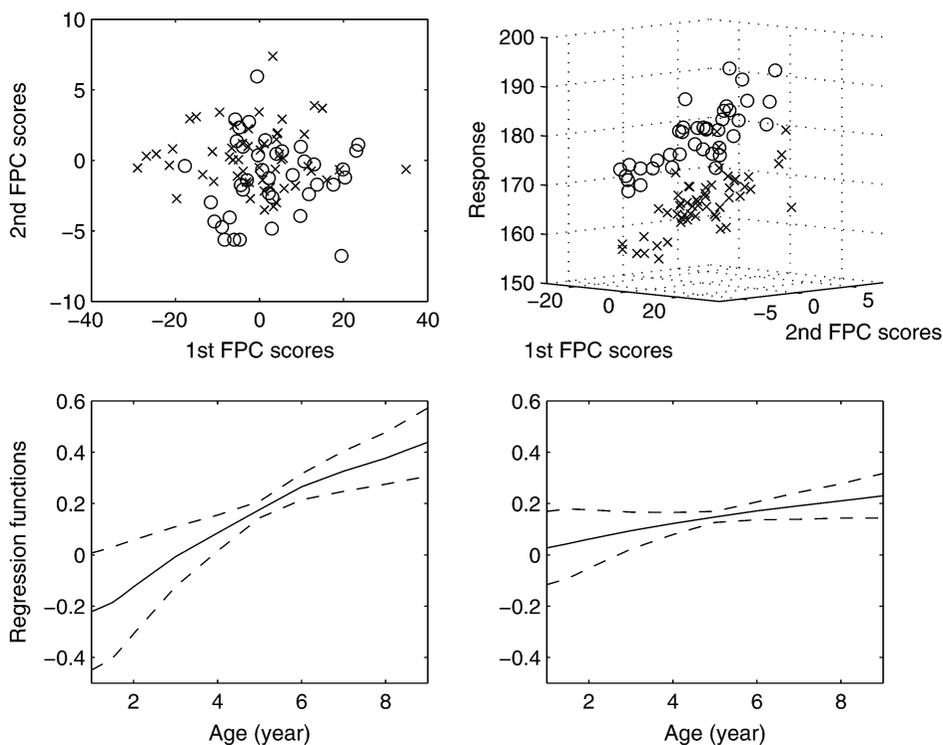


Fig. 5. Top left: the estimated FPC scores obtained by the integral method (2.10) for boys (circles) and girls (crosses). Top right: responses  $Y_i$  (heights at maturity) against estimated FPC scores for boys (circles) and girls (crosses). Bottom panels: estimated regression functions (solid) of 2 groups that correspond to boys (bottom left) and girls (bottom right) along with 95% bootstrap confidence bands (dashed).

increasing trends indicate more weights on the height measurements at later times. It is expected and also confirmed by the data that boys are usually taller than girls at maturity due to the increasing patterns of individual predictor trajectories and the faster ascending regression weights after around age 5. Again by comparison with a FLM based on the CVRPEs, 0.0005 for FMR and 0.0017 for FLM, we observed a substantial reduction of 70%.

#### 4. CONCLUDING REMARKS

In this paper, we investigated a new type of functional regression models, FMR, that relate a scalar response to an infinite-dimensional predictor process through possibly different regression structures. The proposed FMR is particularly useful when the use of a single regression structure for modeling all subjects is inadequate. The need for this modeling approach was demonstrated through 2 real data examples as well as simulation studies that can be found in the online Appendix. Utilizing FPCA as a means for regularization caused by the infinite-dimensional nature of the predictor process, we developed a simple and yet flexible framework that is similar to classical mixture regression with a set of uncorrelated FPC scores as predictors. The estimation procedures can be easily implemented with existing softwares for FPCA and mixture regression. Lastly, we note that the proposed modeling framework can be immediately extended to nonnormal mixture settings and/or to other nonlinear link functions.

#### SUPPLEMENTARY MATERIAL

Supplementary material is available at <http://biostatistics.oxfordjournals.org>.

#### ACKNOWLEDGMENTS

The authors are grateful to the reviewer and the associate editor for many helpful and constructive comments, which led to a significantly improved version of the paper. *Conflict of Interest*: None declared.

#### FUNDING

Discovery Grant from the Natural Sciences and Engineering Research Council of Canada to F.Y., Y.F.; National Science Foundation (0707037 and 1007520) to T.C.M.L.

#### REFERENCES

- CAI, T. AND HALL, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34**, 2159–2179.
- CARDOT, H., FERRATY, F. AND SARDA, P. (1999). Functional linear model. *Statistics and Probability Letters* **45**, 11–22.
- CARDOT, H., FERRATY, F. AND SARDA, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica* **13**, 571–591.
- CAREY, J. R., LIEDO, P., MÜLLER, H. G., WANG, J. L. AND CHIOU, J. M. (1998). Relationship of age patterns of fecundity to mortality, longevity, and lifetime reproduction in a large, cohort of mediterranean fruit fly females. *Journal of Gerontology: Biological Sciences and Medical Sciences* **53**, B245–251.
- CASTRO, P. E., LAWTON, W. H. AND SYLVESTRE, E. A. (1986). Principal modes of variation for processes with continuous sample curves. *Technometrics* **28**, 329–337.
- DESARBO, W. AND CRON, W. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* **5**, 249–282.

- FAN, J., YAO, Q. W. AND CAI, Z. W. (2003). Adaptive varying-coefficient linear models. *Journal of the Royal Statistical Society, Series B* **65**, 57–80.
- FAN, J. AND ZHANG, J. T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B* **62**, 303–322.
- FRALEY, C. AND RAFTERY, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611–631.
- GAFFNEY, S. AND SMYTH, P. (2003). Curve clustering with random effects regression mixtures. In: Bishop, C. and Frey, B. (editors), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, Key West, FL.
- GASSER, T., KÖHLER, W., MÜLLER, H. G., KNEIP, A., LARGO, R., MOLINARI, L. AND PRADER, A. (1984). Velocity and acceleration of height growth using kernel estimation. *Annals of Human Biology* **11**, 397–411.
- HALL, P. AND HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35**, 70–91.
- HENNIG, C. (2000). Identifiability of models for clusterwise linear regression. *Journal of Classification* **17**, 273–296.
- JAMES, L. F., PRIEBE, C. E. AND MARCHETTE, D. J. (2001). Consistent estimation of mixture complexity. *The Annals of Statistics* **29**, 1281–1296.
- JAMES, G. AND SUGAR, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the Royal Statistical Society, Series B* **98**, 397–408.
- JIANG, W. AND TANNER, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics* **27**, 987–1011.
- LUAN, Y. H. AND LI, H. Z. (2003). Clustering of temporal gene expression data using a mixed-effects model with b-splines. *Bioinformatics* **19**, 474–482.
- MORRIS, J. S., VANNUCCI, M., BROWN, P. J. AND CARROLL, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis (with discussion). *Journal of the American Statistical Association* **98**, 573–594.
- MÜLLER, H. G. AND STADTMÜLLER, U. (2005). Generalized functional linear models. *The Annals of Statistics* **33**, 774–805.
- MÜLLER, H. G. AND ZHANG, Y. (2005). Time-varying functional regression for predicting remaining lifetime distributions from longitudinal trajectories. *Biometrics* **61**, 1064–1075.
- PARTRIDGE, L. AND HARVEY, P. H. (1985). Evolutionary biology: cost of reproduction. *Nature* **316**, 20–21.
- RAMSAY, J. O., BOCK, R. D. AND GASSER, T. (1995). Comparisons of heights acceleration curves in the Fels, Zurich and Berkeley growth data. *Annals of Human Biology* **22**, 413–426.
- RAMSAY, J. O. AND SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd edition. New York: Springer.
- RICE, J. AND SILVERMAN, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53**, 233–243.
- TUDDENHAM, R. D. AND SNYDER, M. M. (1954). Physical growth study of California boys and girls from birth to eighteen years. *University of California Publications in Child Development* **1**, 183–364.
- WEDEL, M. AND DESARBO, W. (1995). A mixture likelihood approach for generalized linear models. *Journal of Classification* **12**, 21–55.
- WESTENDORP, R. G. J. AND KIRKWOOD, T. B. L. (1999). Human longevity at the cost of reproductive success. *Nature* **396**, 743–746.

- YAO, F. AND LEE, T. C. M. (2006). Penalized spline models for functional principal component analysis. *Journal of the Royal Statistical Society, Series B* **68**, 3–25.
- YAO, F., MÜLLER, H. G. AND WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

[Received June 17, 2010; revised September 20, 2010; accepted for publication September 21, 2010]

*Biostatistics* (2010), 2, 2, pp. 1–17

doi:10.1093/biostatistics/???

## Supplementary material to functional mixture regression

FANG YAO\*

*Department of Statistics,*

*University of Toronto, Toronto, Ontario M5S 3G3, Canada*

fyao@utstat.toronto.edu

YUEJIAO FU

*Department of Mathematics and Statistics,*

*York University, Toronto, Ontario M3J 1P3, Canada.*

THOMAS C. M. LEE

*Department of Statistics,*

*University of California, Davis, California 95616, U.S.A.*

### 1. SIMULATION STUDIES

We conducted simulation studies in two scenarios to illustrate the empirical performance of the functional mixture regression (FMR) model in terms of both estimation and prediction. We simulated 500 Monte Carlo runs in both scenarios, each run consisting of a collection of  $n = 200$  predictor trajectories  $X_i$  and associated scalar responses  $Y_i$  that serve as the *training sample* for estimation. In addition, for each run, we further gen-

\*To whom correspondence should be addressed.

erated another 200 pairs of  $(X_i, Y_i)$  that constitute the *validation sample*, which will be used towards the end of this section for assessing the predictive power of FMR. All these trajectories were generated with a mean function  $\mu(t) = t + \sin(t)$ ,  $0 \leq t \leq 10$ , and a covariance function derived from two eigenfunctions,  $\phi_1(t) = \sin(\pi t/10)/\sqrt{5}$  and  $\phi_2(t) = \sin(2\pi t/10)/\sqrt{5}$ , associated with eigenvalues  $\lambda_1 = 4$ ,  $\lambda_2 = 1$  as well as  $\lambda_m = 0$  for  $m \geq 3$ . Note that these two eigenfunctions in fact resemble the shapes of the estimated ones in Medfly example. The predictor FPC scores are  $\xi_{im} \sim \mathcal{N}(0, \lambda_m)$ ,  $m = 1, 2$ . The measurement error  $\varepsilon_{ij}$  [(2.9) in the paper] are i.i.d.  $N(0, \sigma_x^2)$ , where two noise levels of the predictor process were considered to demonstrate the influence,  $\sigma_x = 0.1$  and  $0.3$ . Each predictor trajectory was sampled at locations that were uniformly distributed over the domain  $[0, 10]$ . The number of measurements was independently chosen for each trajectory, by selecting a number from  $\{100, \dots, 150\}$  with equal probability.

In Scenario 1 the response was generated from a single regression function,  $\beta(t) = \phi_1(t) + \phi_2(t)$  for  $t \in [0, 10]$ , with an i.i.d. additive noise  $\varepsilon_{i,y}$  distributed as  $N(0, \sigma_y^2)$  for all subjects. We also included two noise levels of the response,  $\sigma_y = 0.2$  and  $0.6$ . In Scenario 2, the response was simulated from two distinct regression functions,  $\beta_1(t) = \phi_1(t) + \phi_2(t)$  for the first 100 subjects and  $\beta_2(t) = \phi_1(t) - \phi_2(t)$  for the rest, and again was contaminated with an i.i.d. additive  $N(0, \sigma_y^2)$  noise  $\varepsilon_{i,y}$ , where  $\sigma_y = 0.2$  and  $\sigma_y = 0.6$  were considered. The proposed FMR was estimated as described in Section 2.3, including automatic choices of various smoothing parameters, the number of FPCs of the predictor processes truncated by the threshold of 90% of overall variation, and the numbers of regression functions chosen by BIC in mixture fitting. It is worth mentioning that  $M = 2$  was correctly specified in most Monte Carlo runs for each case.

We first examine model estimation using the training samples, including the regression coefficients as well as the choice of  $K$ . The benchmark we compared with is the ideal case of fitting the FMR [(2.4) in the paper] using the true FPC scores  $\xi_{im}$ . From Tables 1 and 2, one can see that, for identifying the number of regression functions  $K$ , the proposed FMR methodology is nearly as good as the ideal fitting, where the worst case is 486/500 (97.2%) in the case with the larger noise on predictor process in Scenario 2. These results also provide evidence for the consistency of regression coefficient estimates.

It is of more interest to inspect the predictive ability of the FMR when comparing with the classical functional linear models (FLM). Recall that, for each run, we have generated a validation sample of size  $n = 200$ , and here we use them to calculate the relative prediction error (RPE), defined as  $\text{RPE} = \sum_{i=1}^n (Y_i^* - \hat{Y}_i^*)^2 / \sum_{i=1}^n Y_i^{*2}$ , where  $Y_i^*$  is the response of the  $i$ th new subject in the validation sample and  $\hat{Y}_i^*$  is its predicted value. These predicted values were obtained as follows. First the FPC scores  $\hat{\xi}_{im}^*$  of the new subjects were calculated by applying the integral approach [(2.10) in the paper] to the new noisy predictor trajectories  $U_{ij}^*$ . Then these FPC scores were fed into the fitted FMRs and FLMs respectively to calculate the predicted values, where the parameters of such fitted FMRs and FLMs were estimated from the training sample. It is noticed that the response  $Y_i^*$  were used to determine which cluster the subjects belong to in FMR. From the Monte Carlo estimates of the RPEs listed in Table 3, we see that the FMR achieves dramatic gains ranging from 83% to 91% in Scenario 2, which suggests that the FMR can definitely be a viable alternative when the FLM is not adequate. For Scenario 1, comparable results were obtained for FMR and FLM. This was expected as the true value for  $K = 1$  was correctly specified by FMR in most runs. These comparisons indeed

provide strong evidence for the need of the proposed FMR when a single regression function is not sufficient to characterize the underlying relationship. Also reported in Table 3 are, for Scenario 2, the FMR predictive classification rates for the validation samples that correspond to those runs with  $K$  correctly specified as 2. As expected, they are affected by the noise levels of the predictor process and response.

## 2. THEORETICAL RESULTS

We state in this section the theoretical results on the consistency of the proposed functional mixture regression (FMR) in terms of model estimation and prediction, along with a brief and intuitive outline of the technical arguments. We first need to appropriately quantify the discrepancy between the true and estimated functional principal component (FPC) scores, i.e.,  $\xi_{im}$  and  $\hat{\xi}_{im}$ . Besides needing a large number of subjects, it is also required that the measurements sampled from each subject are sufficiently dense. Then the FPC scores can be satisfactorily estimated by the integral approximation  $\hat{\xi}_{im}^I$  [(2.10) in the paper]. Since the PACE estimates  $\hat{\xi}_{im}^P$  [(2.11) in the paper] can be considered equivalent to  $\hat{\xi}_{im}^I$  in the dense case (Müller, 2005), we shall focus on the integral estimates for theoretical developments and suppress the superscript “ $I$ ” whenever appropriate.

Write  $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iM})^T$  and  $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1}, \dots, \hat{\xi}_{iM})^T$ , where  $M$  is the number of FPCs used for approximation. We call  $\hat{X}_\xi = (\hat{\boldsymbol{\xi}}_1, \dots, \hat{\boldsymbol{\xi}}_n)^T$  the “estimated” design matrix. Given the estimated FPC scores  $\hat{\xi}_{im}$ , any estimate of the parameter  $\boldsymbol{\psi}$  [defined prior to (2.6) in the paper] would in fact be calculated from the “estimated” log-likelihood

$$l_n(\boldsymbol{\psi}; \mathbf{y}, \hat{X}_\xi) = \sum_{i=1}^n l(\boldsymbol{\psi}; y_i, \hat{\boldsymbol{\xi}}_i) = \sum_{i=1}^n \log f(y_i | \hat{\boldsymbol{\xi}}_i, \boldsymbol{\psi}) \quad (2.1)$$

instead of the “true” log-likelihood

$$l_n(\boldsymbol{\psi}; \mathbf{y}, X_\xi) = \sum_{i=1}^n l(\boldsymbol{\psi}; y_i, \boldsymbol{\xi}_i) = \sum_{i=1}^n \log f(y_i | \boldsymbol{\xi}_i, \boldsymbol{\psi}),$$

where  $f(y_i | \boldsymbol{\xi}_i, \boldsymbol{\psi})$  is defined in (2.7) in the paper. Although the consistency of the Maximum Likelihood Estimation (MLE) of  $\boldsymbol{\psi}$  obtained by maximizing the “true” likelihood is applicable to standard mixture regression (Jiang and Tanner, 1999), to the best of our knowledge, there is no existing theory for estimation obtained by maximizing the “estimated” likelihood (2.1). For clarity we denote such an estimate as  $\hat{\boldsymbol{\psi}}$  and call it MLEED, short for MLE based on the Estimated Design matrix  $\hat{X}_\xi$ . A general theorem concerning the consistency of such MLEED has been established in Yao (2010) and is stated in Lemma 3 of Section 4.

We shall consider the case of normal random component and denote the density function of a standard normal by  $\varphi(\cdot)$ . Coupling Lemmas 1–3 in Section 4, together with mild regularity conditions listed in Section 3, we have the following theorem. Recall that  $\Theta$  is the parameter space and  $f(y_i | \boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  is the  $k$ th conditional density, defined in (2.6) and (2.7) in the paper, respectively.

**Theorem 1** Suppose that the assumptions (A1)-(A4) hold with the  $k$ th conditional density  $f(y_i | \boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \varphi\{(y_i - b_{k0} - \boldsymbol{\xi}_i^T \mathbf{b}_k) / \sigma_{ky}\}$ ,  $k = 1, \dots, K$ , and that the true value  $\boldsymbol{\psi}_0$  is an interior point of the parameter space  $\Theta$ . Then, for any compact set  $E \subseteq \Theta$  containing some neighborhood of the true value  $\boldsymbol{\psi}_0$ , there exists a sequence of estimates  $\hat{\boldsymbol{\psi}} \equiv \hat{\boldsymbol{\psi}}_n$  maximizing the estimated likelihood function  $l_n(\boldsymbol{\psi}; \mathbf{y}, \hat{X}_\xi)$  on  $E$ , such that  $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}_0$ .

Our estimates aim for the regression parameter functions  $\beta_{k,M}(t) = \sum_{m=1}^M b_{km} \phi_m(t)$

[(2.5) in the paper],  $k = 1, \dots, K$ . Let  $b_{km}^{(0)}$ ,  $\beta_{k,M}^{(0)}(t)$  and  $E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k)$  [(2.4) in the paper] be the quantities evaluated at the true values  $\boldsymbol{\psi}_0$ ,  $\phi_m$  and  $\xi_{im}$ . That is,  $\beta_{k,M}^{(0)}(t) = \sum_{m=1}^M b_{km}^{(0)}\phi_m(t)$  and  $E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k) = b_{k0}^{(0)} + \sum_{m=1}^M b_{km}^{(0)}\xi_{im}$ , where  $t \in \mathcal{T}$ ,  $k = 1, \dots, K$ ,  $i = 1, \dots, n$ . Then we can obtain consistent estimation and prediction both individually and on average.

**Theorem 2** If the assumptions in Theorem 1 hold, for any compact set  $E \subseteq \Theta$  containing some neighborhood of  $\boldsymbol{\psi}_0$ , letting  $\hat{\beta}_{k,M}(t)$  and  $\hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k)$  be the quantities evaluated at  $\hat{\phi}_m$ ,  $\hat{\xi}_{im}$  and  $\hat{\boldsymbol{\psi}}$  that maximizes  $l_n(\boldsymbol{\psi}; \mathbf{y}, \hat{X}_\xi)$  on  $E$ , i.e.,  $\hat{\beta}_{k,M}(t) = \sum_{m=1}^M \hat{b}_{km}\hat{\phi}_m(t)$  and  $\hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k) = \hat{b}_{k0} + \sum_{m=1}^M \hat{b}_{km}\hat{\xi}_{im}$ , then

$$\sup_{t \in \mathcal{T}} |\hat{\beta}_{k,M}(t) - \beta_{k,M}^{(0)}(t)| \xrightarrow{p} 0, \quad \text{for } k = 1, \dots, K, \quad (2.2)$$

$$\hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k) - E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k) \xrightarrow{p} 0, \quad \text{for } i = 1, \dots, n, \quad (2.3)$$

$$\frac{1}{n} \sum_{i=1}^n \{ \hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k) - E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k) \} \xrightarrow{p} 0. \quad (2.4)$$

*Remark.* In principle, the consistency of MLEED  $\hat{\boldsymbol{\psi}}$  as well as the predictions can be extended to FMR model with other conditional densities  $f(y_i|\boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  and/or with suitable nonlinear link functions  $g(b_{k0} + \boldsymbol{\xi}_i^T \mathbf{b}_k)$ , provided that the conditions in Lemma 3 and other necessary regularity conditions are fulfilled.

### 3. TECHNICAL ASSUMPTIONS

Necessary assumptions are listed below. Briefly, these assumptions concern the number and density of measurements per trajectory, the underlying stochastic process  $X(t)$  and the noise process  $U(t)$  that generates the observed repeated measurements  $U_{ij}$  [(2.9) in the paper], as well as various smoothing parameters and kernel functions. Let  $b = b(n)$ ,

$h = h(n)$  and  $h^* = h^*(n)$  denote the bandwidths for estimating  $\hat{\mu}$  (26),  $\hat{G}$  (27) and  $\hat{\sigma}_x$  (2) in Yao, Müller and Wang (2005).

$$(A1) \quad b \rightarrow 0, h^* \rightarrow 0, h \rightarrow 0, nb^2 \rightarrow \infty, nh^{*2} \rightarrow \infty, nh^4 \rightarrow \infty, nb^6 < \infty, \\ nh^{*6} < \infty, nh^8 < \infty, \text{ as } n \rightarrow \infty,$$

Denote the sorted time points across all subjects as  $a_0 \leq t_{(1)} \leq \dots \leq t_{(N_n)} \leq b_0$ , and  $\Delta = \max\{t_{(k)} - t_{(k-1)} : k = 1, \dots, N+1\}$ , where  $N_n = \sum_{i=1}^n n_i$ ,  $\mathcal{T} = [a_0, b_0]$ ,  $t_{(0)} = a_0$ , and  $t_{(N+1)} = b_0$ . For the  $i$ th subject, suppose that the time points  $t_{ij}$  have been ordered non-decreasingly. Let  $\Delta_i = \max\{t_{ij} - t_{i,j-1} : j = 1, \dots, n_i + 1\}$  and  $\Delta^* = \max\{\Delta_i : i = 1, \dots, n\}$ , where  $t_{i0} = a_0$  and  $t_{i,n_i+1} = b_0$ , and  $\bar{n} = n^{-1} \sum_{i=1}^n n_i$ . To obtain consistent FPC score estimates, we require both the pooled data across all subjects and the data from each subject to be dense in the time domain  $\mathcal{T}$ . For convenience, we study the asymptotics in the manner of  $\bar{n} \rightarrow \infty$  as  $n \rightarrow \infty$ , and assume that

$$(A2) \quad \Delta = O(\min\{n^{-1/2}b^{-1}, n^{-1/2}h^{*-1}, n^{-1/4}h^{-1}\}), \max\{n_i : i = 1, \dots, n\} \leq \\ C\bar{n} \text{ for some } C > 0, \text{ and } \Delta^* = O(1/\bar{n}), \text{ as } n \rightarrow \infty.$$

Denote by  $U_i(t) \stackrel{\text{i.i.d.}}{\sim} U(t)$  the distribution that generates  $U_{ij}$  for the  $i$ th subject at  $t_{ij}$ . The predictor process  $X$  and measurement  $U$  are assumed to satisfy the following conditions.

$$(A3) \quad E(\|X'\|_\infty^2) < \infty, E(\|X'^2\|_\infty^2) = o(\bar{n}_x), \sup_{t \in \mathcal{T}} E[U^4(t)] < \infty.$$

Recall that smoothing kernels  $K_1$  and  $K_2$  are compactly supported densities with zero means and finite variances. The Fourier transformations of  $K_1$  and  $K_2$  are denoted by  $\kappa_1(t) = \int e^{-iut} K_1(u) du$  and  $\kappa_2(t, s) = \int e^{-(iut+ivs)} K_2(u, v) du dv$  respectively. We require

(A4)  $\int |\kappa_1(t)|dt < \infty$ ,  $\int \int |\kappa_2(t, s)|dtds < \infty$ , i.e.,  $\kappa_1(t)$  and  $\kappa_2(t, s)$  are both absolutely integrable.

Let  $g_1(u; t)$  denote the density function of  $U(t)$ , and  $g_2(u_1, u_2; t_1, t_2)$  denote the density of  $(U(t_1), U(t_2))$ . It is assumed throughout that these density functions satisfy appropriate regularity conditions.

#### 4. AUXILIARY LEMMAS

Denote the true and estimated covariance operators by  $\mathbf{G}$  and  $\widehat{\mathbf{G}}$ , generated by  $G$  and  $\widehat{G}$  respectively; i.e.,  $\mathbf{G}(f) = \int_{\mathcal{T}} G(s, t)f(s)ds$  and  $\widehat{\mathbf{G}}(f) = \int_{\mathcal{T}} \widehat{G}(s, t)f(s)ds$  for any  $f \in L^2(\mathcal{T})$ . Define

$$\begin{aligned} D_X &= [\int_{\mathcal{T}^2} \{\widehat{G}(s, t) - G(s, t)\}^2 dsdt]^{1/2}, & \delta_m &= \min_{1 \leq j \leq m} (\lambda_j - \lambda_{j+1}), \\ M^* &= \inf\{j \geq 1 : \lambda_j - \lambda_{j+1} \leq 2D_X\} - 1, & \pi_m &= 1/\lambda_m + 1/\delta_m. \end{aligned} \quad (4.1)$$

**Lemma 1** Under (A1)-(A4) and appropriate regularity conditions for density functions  $g_1(u, t)$  and  $g_2(u_1, u_2; t_1, t_2)$ ,

$$\sup_{t \in \mathcal{T}} |\hat{\mu}(t) - \mu(t)| = O_p\left(\frac{1}{\sqrt{nb}}\right), \quad \sup_{s, t \in \mathcal{T}} |\widehat{G}(s, t) - G(s, t)| = O_p\left(\frac{1}{\sqrt{nh^2}}\right), \quad (4.2)$$

and as a consequence,  $\hat{\sigma}_x^2 - \sigma_x^2 = O_p(n^{-1/2}h^{-2} + n^{-1/2}h^{*-1})$ . Considering eigenvalues  $\lambda_m$  of multiplicity one,  $\hat{\phi}_m$  can be chosen such that,  $m = 1, \dots, M^*$ ,

$$P\left(\sup_{1 \leq m \leq M} |\hat{\lambda}_m - \lambda_m| \leq D_X\right) = 1, \quad \sup_{t \in \mathcal{T}} |\hat{\phi}_m(t) - \phi_m(t)| = O_p\left(\frac{\pi_m}{\sqrt{nh^2}}\right), \quad (4.3)$$

where  $D_X$ ,  $\pi_m$  and  $M^*$  are defined in (4.1).

The next lemma provides upper bounds for the estimation errors  $|\hat{\xi}_{im}^I - \xi_{im}|$  with some specific structure, and the derivation can be found in Müller and Yao (2008). Let  $\|f\|_\infty = \sup_{t \in \mathcal{A}} |f(t)|$  for an arbitrary function  $f$  with support  $\mathcal{A}$ ,  $\|g\| = \sqrt{\int_{\mathcal{A}} g^2(t) dt}$  for any  $g \in L^2(\mathcal{A})$ , and define

$$\begin{aligned}
\theta_{im}^{(1)} &= c_1 \|X_i\| + c_2 \|X_i X_i'\|_\infty \Delta^* + c_3, & Z_m^{(1)} &= \sup_{t \in \mathcal{T}} |\hat{\phi}_m(t) - \phi_m(t)|, \\
\theta_{im}^{(2)} &= 1 + \|\phi_m \phi_m'\|_\infty \Delta^*, & Z_m^{(2)} &= \sup_{t \in \mathcal{T}} |\hat{\mu}(t) - \mu(t)|, \\
\theta_{im}^{(3)} &= c_4 \|X_i\|_\infty + c_5 \|X_i'\|_\infty + c_6, & Z_m^{(3)} &= \|\phi_m'\|_\infty \Delta^*, \\
\theta_{im}^{(4)} &= \left| \sum_{j=2}^{n_i} \varepsilon_{ij} \phi_m(t_{ij})(t_{ij} - t_{i,j-1}) \right|, & Z_m^{(4)} &\equiv 1, \\
\theta_{im}^{(5)} &= \sum_{j=2}^{n_i} |\varepsilon_{ij}| (t_{ij} - t_{i,j-1}), & Z_m^{(5)} &\equiv Z_m^{(1)},
\end{aligned} \tag{4.4}$$

for some positive constants  $c_1, \dots, c_6$  that do not depend on  $i$  or  $m$ . We note that the subscripts are mainly for notational convenience and do not necessarily reflect dependence on these indices. More importantly, we emphasize that  $\theta_{im}^{(\ell)}$  are i.i.d. over  $i$  ( $\ell = 1, 3, 4, 5$ ) or nonrandom that is free  $i$  ( $\ell = 2$ ), and that the  $Z_m^{(\ell)}$  do not depend on  $i$  for all  $\ell = 1, 2, 3, 4, 5$ .

**Lemma 2** For integral estimates of the FPC scores  $\hat{\xi}_{im}^I$  in (2.10) of the paper,

$$|\hat{\xi}_{im}^I - \xi_{im}| \leq \sum_{\ell=1}^5 \theta_{im}^{(\ell)} Z_m^{(\ell)}, \quad m = 1, \dots, M^*, \tag{4.5}$$

where  $\theta_{im}^{(\ell)}$  and  $Z_m^{(\ell)}$  are defined in (4.4), and  $M^*$  is defined in (4.1).

We aim for the consistency results for any  $M$  and  $K$ , where  $M$  and  $K$  are the numbers of FPCs and distinct regression functions in the FMR model. We state a useful theorem proved in Yao (2010) as Lemma 3 regarding the consistency of the Maximum Likelihood

Estimation based on Estimated Design (MLEED), that can be shown applicable to the proposed FMR. For convenience, we first define some conditions that are required for some relevant functions. A function  $h(\boldsymbol{\psi}, y, \boldsymbol{\xi})$  is said to satisfy the assumption (B1) at  $\boldsymbol{\psi}_1 \in \Theta$ , provided that the following holds.

(B1) There exist some functions  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  and  $c(\boldsymbol{\psi})$  such that, for all possible values of  $y, \boldsymbol{\xi}', \boldsymbol{\xi}''$ , and  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$ , where  $N_{\boldsymbol{\psi}_1}$  is some neighborhood of  $\boldsymbol{\psi}_1$ ,

$$\|h(\boldsymbol{\psi}, y, \boldsymbol{\xi}'') - h(\boldsymbol{\psi}, y, \boldsymbol{\xi}')\| \leq g(y, \boldsymbol{\xi}', \boldsymbol{\psi}) \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\| + c(\boldsymbol{\psi}) \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\|^2,$$

and  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  and  $c(\boldsymbol{\psi})$  satisfy

$$\sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} E_{(\boldsymbol{\psi}_0, \Lambda_0)} \{g^2(Y, \boldsymbol{\xi}, \boldsymbol{\psi})\} < \infty, \quad \sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} c(\boldsymbol{\psi}) < \infty,$$

where  $\boldsymbol{\psi}_0$  and  $\Lambda_0$  are the true values of  $\boldsymbol{\psi}$  and  $\Lambda$ .

A function  $q(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  is said to satisfy the set of assumptions (B2) at  $\boldsymbol{\psi}_1 \in \Theta$ , if the conditions (B2.1)–(B2.3) below hold.

(B2.1)  $q(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  is upper semicontinuous in  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$  for all  $(y, \boldsymbol{\xi})$ ;

(B2.2) There exists a function  $D(y, \boldsymbol{\xi})$  such that  $E_{(\boldsymbol{\psi}_0, \Lambda_0)} D(y, \boldsymbol{\xi}) < \infty$  and  $q(y, \boldsymbol{\xi}, \boldsymbol{\psi}) \leq D(y, \boldsymbol{\xi})$  for all  $(y, \boldsymbol{\xi})$  and  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$ ;

(B2.3) For  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$  and sufficiently small  $r > 0$ ,  $\sup_{\{\boldsymbol{\psi}': \|\boldsymbol{\psi}' - \boldsymbol{\psi}\| < r\}} q(y, \boldsymbol{\xi}, \boldsymbol{\psi}')$  is measurable in  $(y, \boldsymbol{\xi})$ .

In Lemma 3, let  $f(y|\boldsymbol{\xi}, \boldsymbol{\psi})$ ,  $\boldsymbol{\psi} \in \Theta$  denote a general conditional density function with a parameter space  $\Theta$  that is a subset of  $\mathcal{R}^p$  for some positive integer  $p$  [not restricted to the

conditional density defined in (9) of Section 2.2], and  $\hat{\xi}_i$  be any sequence of estimates of  $\xi_i$ ,  $i = 1, \dots, n$ . Denote  $l(\boldsymbol{\psi}; y, \boldsymbol{\xi}) = \log f(y|\boldsymbol{\xi}, \boldsymbol{\psi})$ .

**Lemma 3** Suppose that the true value  $\boldsymbol{\psi}_0$  is an interior point of the parameter space  $\Theta$ . Consider an arbitrary compact set  $E$  satisfying  $N_{\boldsymbol{\psi}_0} \subseteq E \subseteq \Theta$  and any  $\boldsymbol{\psi} \in E$  is an interior point of  $\Theta$ , where  $N_{\boldsymbol{\psi}_0}$  is some neighborhood of  $\boldsymbol{\psi}_0$ . Assume that

- (i) There exist some  $Z_n^{(j)}$  and  $\theta_{i,n}^{(j)}$ , where  $\theta_{i,n}^{(j)}$  are either i.i.d. realizations of positive random variables  $\theta_n^{(j)}$  or nonrandom constants with respect to  $i$ , where  $j = 1, \dots, J$ , for some finite  $J$ ,

$$\|\hat{\xi}_i - \xi_i\| \leq \sum_{j=1}^J Z_n^{(j)} \theta_{i,n}^{(j)}, \quad E\{(\theta_n^{(j)})^2\} < \infty, \quad Z_n^{(j)} \sqrt{E\{(\theta_n^{(j)})^2\}} \xrightarrow{p} 0;$$

- (ii) For any  $\boldsymbol{\psi} \in E$ ,  $l(\boldsymbol{\psi}; y, \boldsymbol{\xi})$  satisfies the assumptions (B1) at  $\boldsymbol{\psi}$  with some functions  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  and  $c(\boldsymbol{\psi})$ , where  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  satisfies the assumptions (B2) at  $\boldsymbol{\psi}$ ;
- (iii) For any  $\boldsymbol{\psi} \in E$ ,  $l(\boldsymbol{\psi}; y, \boldsymbol{\xi})$  satisfies the assumptions (B2) at  $\boldsymbol{\psi}$ ;
- (iv)  $f(y|\boldsymbol{\xi}, \boldsymbol{\psi}) = f(y|\boldsymbol{\xi}, \boldsymbol{\psi}^*)$  implies that  $\boldsymbol{\psi} = \boldsymbol{\psi}^*$  in a well-defined sense (identifiability).

Then for any sequence of the maximizer  $\hat{\boldsymbol{\psi}}$  of  $l_n(\boldsymbol{\psi}; y, \hat{X}_\xi) = \sum_{i=1}^n l(\boldsymbol{\psi}; y_i, \hat{\xi}_i)$  on the compact set  $E$ , i.e., the maximum likelihood estimates based on estimated design (MLEED), one has

$$\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}_0.$$

## 5. PROOFS OF MAIN THEOREMS

*Proof of Theorem 1.* We first verify the conditions (i)–(iv) in Lemma 3 for the FMR model with the conditional density of the  $k$ th component  $f(y_i|\boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \varphi\{(y_i - b_{k0} - \boldsymbol{\xi}_i^T \mathbf{b}_k)/\sigma_{ky}\}$ ,  $k = 1, \dots, K$ , where  $\varphi(u) = \exp(-u^2/2)/\sqrt{2\pi}$  is the density function of standard normal. Given the expressions of  $Z_m^{(\ell)}$  and  $\theta_{im}^{(\ell)}$  in (4.4), one notes that  $\theta_{im}^{(\ell)}$  are i.i.d. or nonrandom w.r.t.  $i$ ,  $\ell = 1, \dots, 5$ ,  $m = 1, \dots, M$ . With (A3), it is obvious that  $E\{(\theta_{im}^{(\ell)})^2\} < \infty$  for  $\ell = 1, 2, 3$ . Due to the orthonormality of  $\phi_m$  and the independence among  $\varepsilon_{ij}$ 's, we have  $E\{(\theta_{im}^{(4)})^2\} = E\{[\sum_{j=2}^{n_i} \varepsilon_{ij} \phi_m(t_{ij})(t_{ij} - t_{i,j-1})]^2\} = \text{var}\{\sum_{j=2}^{n_i} \varepsilon_{ij} \phi_m(t_{ij})(t_{ij} - t_{i,j-1})\} = \sigma_x^2 \sum_{j=2}^{n_i} \phi_m^2(t_{ij})(t_{ij} - t_{i,j-1})^2 \leq 2\sigma^2 \Delta^* \rightarrow 0$ . For  $\theta_{im}^{(5)}$ , applying Cauchy-Schwartz inequality,  $E\{(\theta_{im}^{(4)})^2\} \leq \{\sum_{j=2}^{n_i} E(\varepsilon_{ij}^2)(t_{ij} - t_{i,j-1})\} \mathcal{T} \leq 2\mathcal{T}^2 \sigma_x^2 < \infty$  for large  $n$ . Combining with Lemmas 1 and 2, then condition (i) holds. Since the parameter space  $\Theta$  defined is an open subset of  $\mathcal{R}^{(M+3)K-1}$ , any  $\boldsymbol{\psi} \in E$  is always an interior point of  $\Theta$ . It is easy to verify that condition (iii) holds for the conditional density  $f(y_i|\boldsymbol{\xi}_i, \boldsymbol{\psi})$  with normal components, while condition (iv) is satisfied given the identifiability in the sense of (2.8) in the paper.

Now we check condition (ii), and observe that

$$l(\boldsymbol{\psi}; y, \boldsymbol{\xi}) = \log \left\{ \sum_{k=1}^K \pi_k f(y|\boldsymbol{\xi}, b_{k0}, \mathbf{b}_k, \sigma_{ky}) \right\},$$

$$f(y|\boldsymbol{\xi}, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ky}} \exp \left\{ -\frac{(y - b_{k0} - \boldsymbol{\xi}^T \mathbf{b}_k)^2}{2\sigma_{ky}^2} \right\}. \quad (5.1)$$

For any fixed interior point  $\boldsymbol{\psi}_1$  of  $\Theta$ , one can always assume that a sufficiently small neighborhood  $N_{\boldsymbol{\psi}_1}$  is bounded, and particularly  $\delta \leq \pi_k \leq 1 - \delta$  and  $\sigma_{ky} > \delta$  for some

$\delta > 0, k = 1, \dots, K$ . By Mean Value Theorem, one has, for  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$ ,

$$\begin{aligned} l(\boldsymbol{\psi}; y, \boldsymbol{\xi}'') - l(\boldsymbol{\psi}; y, \boldsymbol{\xi}') &= (\partial^T l(\boldsymbol{\psi}; y, \boldsymbol{\xi}^*) / \partial \boldsymbol{\xi})(\boldsymbol{\xi}'' - \boldsymbol{\xi}'), \\ \frac{\partial}{\partial \boldsymbol{\xi}} l(\boldsymbol{\psi}; y, \boldsymbol{\xi}^*) &= \frac{\sum_{k=1}^K \pi_k f(y | \boldsymbol{\xi}^*, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) (y - b_{k0} - \boldsymbol{\xi}^{*T} \mathbf{b}_k) \mathbf{b}_k / \sigma_{ky}^2}{\sum_{k=1}^K \pi_k f(y | \boldsymbol{\xi}^*, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)}, \end{aligned}$$

where  $\boldsymbol{\xi}^* = \boldsymbol{\xi}' + v(\boldsymbol{\xi}'' - \boldsymbol{\xi}')$  for some  $0 \leq v \leq 1$ . In spite of the complex appearance of the above expression, one can see that it is in fact a *weighted average* of  $(y - b_{k0} - \boldsymbol{\xi}^{*T} \mathbf{b}_k) \mathbf{b}_k / \sigma_{ky}^2$  with weights  $\pi_k f(y | \boldsymbol{\xi}^*, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$ ,  $k = 1, \dots, K$ . Therefore,

$$\begin{aligned} \left\| \frac{\partial}{\partial \boldsymbol{\xi}} l(\boldsymbol{\psi}; y, \boldsymbol{\xi}^*) \right\| &\leq \sum_{k=1}^K \left\| (y - b_{k0} - \boldsymbol{\xi}^{*T} \mathbf{b}_k) \mathbf{b}_k / \sigma_{ky}^2 \right\| \\ &\leq \sum_{k=1}^K \frac{1}{\sigma_{ky}^2} \left\{ \|\mathbf{b}_k y - b_{k0} \mathbf{b}_k\| + \|\mathbf{b}_k\|^2 \|\boldsymbol{\xi}' + v(\boldsymbol{\xi}'' - \boldsymbol{\xi}')\| \right\} \\ &\leq \sum_{k=1}^K \frac{\|\mathbf{b}_k\|}{\sigma_{ky}^2} \left\{ |y - b_{k0}| + \|\mathbf{b}_k\| \|\boldsymbol{\xi}'\| \right\} + \left\{ \sum_{k=1}^K \frac{\|\mathbf{b}_k\|^2}{\sigma_{ky}^2} \right\} \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\| \\ &\equiv g(y, \boldsymbol{\xi}', \boldsymbol{\psi}) + c(\boldsymbol{\psi}) \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\|. \end{aligned}$$

From the boundedness of the small  $N_{\boldsymbol{\psi}_1}$ , it is easy to see that  $\sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} c(\boldsymbol{\psi}) < \infty$ ,  $\sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} E_{(\boldsymbol{\psi}_0, \Lambda_0)} \{g^2(Y, \boldsymbol{\xi}, \boldsymbol{\psi})\} < \infty$ , and moreover  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  satisfies the assumptions (B2) at  $\boldsymbol{\psi}_1$ . Thus condition (ii) holds. The existence of a consistent sequence  $\hat{\boldsymbol{\psi}} \in E$  that are roots of  $\partial l_n(\boldsymbol{\psi}; y, \hat{\boldsymbol{\xi}}) / \partial \boldsymbol{\psi} = 0$  follows for the conditional mixture normal density (5.1).

*Proof of Theorem 2.* The uniform consistency of  $\hat{\beta}_{k,M}(t)$  w.r.t.  $t \in \mathcal{T}$  is obvious given Theorem 1 and Lemma 1. For individual prediction, note that  $|\hat{E}(Y_i | X_i, M) - E(Y_i | X_i, M)| \leq |b_{k0} - b_{k0}| + \|\mathbf{b}_k - \mathbf{b}_k\| \cdot \|\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\|$  and  $\|\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\| \leq \sum_{m=1}^M \sum_{\ell=1}^5 Z_m^{(\ell)} \theta_{im}^{(\ell)}$ .

We have shown that  $E\{(\theta_m^{(\ell)})^2\} < \infty$  and  $Z_m^{(\ell)} \sqrt{E\{(\theta_m^{(\ell)})^2\}} \xrightarrow{p} 0$ , where  $\theta_{im}^{(\ell)} \stackrel{\text{i.i.d.}}{\sim} \theta_m^{(\ell)}$  (considering i.i.d. random variables here without loss of generality),  $\ell = 1, \dots, 5$ ,  $m = 1, \dots, M$ . We then arrive at the result (2.3) by observing the following for each  $m$  and  $\ell$ . For any  $\epsilon > 0$  and  $\delta > 0$ , we choose  $A \geq \sqrt{2/(\epsilon\delta)}$ , i.e.,  $(A^2\delta)^{-1} \leq \epsilon/2$ , when  $n$  is sufficiently large, and apply Chebyshev's inequality:

$$\begin{aligned} P\left(Z_m^{(\ell)}|\theta_{im}^{(\ell)} - E\theta_m^{(\ell)}| > \delta\right) &\leq P\left(Z_m^{(\ell)}\sqrt{E\{(\theta_m^{(\ell)})^2\}} > \frac{\sqrt{\delta}}{A}\right) + P\left(\frac{|\theta_{im}^{(\ell)} - E\theta_m^{(\ell)}|}{A\sqrt{E\{(\theta_m^{(\ell)})^2\}}} > \sqrt{\delta}\right) \\ &\leq \frac{\epsilon}{2} + \frac{1}{A^2\delta} \leq \epsilon. \end{aligned}$$

Noting that  $(1/n)\|\hat{\xi}_i - \xi_i\| \leq \sum_{m=1}^M \sum_{\ell=1}^5 Z_m^{(\ell)}(1/n) \sum_{i=1}^n \theta_{im}^{(\ell)}$ , then the consistency of the average prediction (2.4) follows immediately from the law of large numbers for triangular arrays.

## 6. EM ALGORITHM FOR MIXTURE REGRESSION MODELS

For completeness we outline an EM algorithm for fitting mixture regression models. For details, see, for example, Naik, Shi and Tsai (2007).

Consider the following mixture model with  $K$  normal density components:

$$f(y_i|\xi_i, \psi) = \sum_{k=1}^K \pi_k \varphi(y_i|\xi_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2), \quad i = 1, \dots, n,$$

where  $0 < \pi_k < 1$  and  $\sum \pi_k = 1$ ,  $\varphi(y_i|\xi_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  is the normal density with mean  $(b_{k0} + \xi_i^T \mathbf{b}_k)$  and variance  $\sigma_{ky}^2$ , and  $\psi = (\mathbf{b}_0^T, \mathbf{b}_1^T, \dots, \mathbf{b}_K^T, \boldsymbol{\pi}^T, \boldsymbol{\sigma}_y^T)^T$  is the vector containing all relevant parameters. Let  $\psi^{(r)}$  be the  $r$ -th iterative estimate for  $\psi$ . In the

E-Step of the algorithm, one calculates

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \varphi(y_i | \boldsymbol{\xi}_i, b_{k0}^{(r)}, \mathbf{b}_k^{(r)}, \sigma_{ky}^{2(r)})}{\sum_{k=1}^K \pi_k^{(r)} \varphi(y_i | \boldsymbol{\xi}_i, b_{k0}^{(r)}, \mathbf{b}_k^{(r)}, \sigma_{ky}^{2(r)})}.$$

This quantity can be seen as the  $r$ -th estimated probability for  $y_i$  originated from the  $k$ -th component.

In the M-Step, the  $(r+1)$ -th estimates are calculated with the following closed-form expressions:  $\pi_k^{(r+1)} = n^{-1} \sum_{i=1}^n \tau_{ik}^{(r)}$  and

$$\begin{pmatrix} b_{k0}^{(r+1)} \\ \mathbf{b}_k^{(r+1)} \end{pmatrix} = (\tilde{X}_k^{(r)T} \tilde{X}_k^{(r)})^{-1} \tilde{X}_k^{(r)T} \tilde{\mathbf{y}}_k^{(r)}, \quad \sigma_{ky}^{2(r+1)} = \frac{\tilde{\mathbf{y}}_k^{(r)T} (I - \tilde{H}_k^{(r)}) \tilde{\mathbf{y}}_k^{(r)}}{\text{tr}(W_k^{(r)})},$$

for  $k = 1, \dots, K$ . In the above  $W_k^{(r)} = \text{diag}(\tau_{1k}^{(r)}, \dots, \tau_{nk}^{(r)})$ ,  $\tilde{X}_k^{(r)} = W_k^{(r)1/2} \tilde{X}_\xi$ ,  $\tilde{\mathbf{y}}_k^{(r)} = W_k^{(r)1/2} \mathbf{y}$ ,  $\tilde{H}_k^{(r)} = \tilde{X}_k^{(r)} (\tilde{X}_k^{(r)T} \tilde{X}_k^{(r)})^{-1} \tilde{X}_k^{(r)T}$ ,  $\tilde{X}_\xi = (\mathbf{1}, X_\xi)$ ,  $X_\xi = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n)^T$  and  $\mathbf{y} = (y_1, \dots, y_n)^T$ .

## REFERENCES

- JIANG, W. AND TANNER, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Annals of Statistics* **27**, 987–1011.
- MÜLLER, H. G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32**, 223–240.
- MÜLLER, H. G. AND YAO, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.
- NAIK, P. A., SHI, P. AND TSAI, C. L. (2007). Extending the Akaike Information Criterion to Mixture Regression Models. *Journal of the American Statistical Association* **102**, 244–254.
- YAO, F. (2010). Maximum likelihood estimation based on estimated design or data. *Technical Report*.

YAO, F., MÜLLER, H. G. AND WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

Table 1. *Section 3 Simulation Scenario 1. Monte Carlo estimates of regression coefficients (standard errors in parentheses) for 4 combinations of noise levels, calculated from those runs that  $K = 1$  were correctly specified. The first integer in each case reports the number, out of 500 runs, of correctly specified runs. In this scenario there is one single regression function with true values  $b_{11} = b_{12} = 1$ . The first row corresponds to the ideal fitting (IDEAL) while the second row corresponds to FMR.*

		Noises	$\sigma_x = .1$		$\sigma_x = .3$	
IDEAL	$\sigma_y = .2$	498	.9996	.9998	496	.9998 1.0002
FMR		495	.9796	.9975	494	.9179 .9982
			(.0073)	(.0152)		(.0049) (.0070)
			(.0493)	(.0546)		(.0798) (.0770)
IDEAL	$\sigma_y = .6$	497	1.0002	1.0004	496	1.0002 1.0009
FMR		497	.9908	.9934	497	.9251 .9942
			(.0141)	(.0206)		(.0147) (.0216)
			(.0861)	(.0867)		(.0835) (.0854)

Table 2. *Similar to Table 1 but for Scenario 2. For this scenario the true value for  $K = 2$  and the regression coefficients are  $(b_{11}, b_{12}, b_{21}, b_{22}) = (1, 1, 1, -1)$ .*

		Noises	$\sigma_x = .1$				$\sigma_x = .3$				
IDEAL	$\sigma_y = .2$	494	.9998	.9999	.9999	-1.0007	495	.9998	1.0004	.9997	-.9999
FMR		494	.9786	.9992	.9780	-1.0014	486	.9969	.8422	1.0114	-.8211
			(.0111)	(.0217)	(.0110)	(.0208)		(.0067)	(.0100)	(.0072)	(.0101)
			(.0519)	(.0628)	(.0531)	(.0603)		(.0832)	(.0838)	(.0827)	(.0839)
IDEAL	$\sigma_y = .6$	493	1.0004	1.0008	1.0009	-1.0009	496	1.0009	.9988	1.0001	-.9980
FMR		492	.9922	.9937	.9842	-1.0013	488	.9261	.9933	.9197	-.9979
			(.02189)	(.0293)	(.0235)	(.0315)		(.0222)	(.0317)	(.0225)	(.0310)
			(.0895)	(.0880)	(.0887)	(.0922)		(.0913)	(.0954)	(.0945)	(.0953)

Table 3. Monte Carlo estimates of the relative prediction errors (RPE) defined in Section 3 for 4 combinations of noise levels. Also reported in the last row is the predictive classification rates (P. C. Rate) calculated for the validation samples that correspond to those runs with  $K = 2$  correctly specified.

Model	Method	Noise levels: $\{\sigma_x, \sigma_y\}$			
		$\{.1, .2\}$	$\{.1, .6\}$	$\{.3, .2\}$	$\{.3, .6\}$
Scenario I ( $K = 1$ )	FLM	.02448	.03408	.05764	.08152
	FMR	.02447	.03408	.05764	.08152
Scenario II ( $K = 2$ )	FLM	.23210	.34332	.37007	.39004
	FMR	.0218	.03016	.04943	.06804
	(P. C. Rate)	(.8932)	(.8951)	(.8664)	(.8448)

*Biostatistics* (2010), 2, 2, pp. 1–17

doi:10.1093/biostatistics/???

## Supplementary material to functional mixture regression

FANG YAO\*

*Department of Statistics,*

*University of Toronto, Toronto, Ontario M5S 3G3, Canada*

fyao@utstat.toronto.edu

YUEJIAO FU

*Department of Mathematics and Statistics,*

*York University, Toronto, Ontario M3J 1P3, Canada.*

THOMAS C. M. LEE

*Department of Statistics,*

*University of California, Davis, California 95616, U.S.A.*

### 1. SIMULATION STUDIES

We conducted simulation studies in two scenarios to illustrate the empirical performance of the functional mixture regression (FMR) model in terms of both estimation and prediction. We simulated 500 Monte Carlo runs in both scenarios, each run consisting of a collection of  $n = 200$  predictor trajectories  $X_i$  and associated scalar responses  $Y_i$  that serve as the *training sample* for estimation. In addition, for each run, we further gen-

\*To whom correspondence should be addressed.

erated another 200 pairs of  $(X_i, Y_i)$  that constitute the *validation sample*, which will be used towards the end of this section for assessing the predictive power of FMR. All these trajectories were generated with a mean function  $\mu(t) = t + \sin(t)$ ,  $0 \leq t \leq 10$ , and a covariance function derived from two eigenfunctions,  $\phi_1(t) = \sin(\pi t/10)/\sqrt{5}$  and  $\phi_2(t) = \sin(2\pi t/10)/\sqrt{5}$ , associated with eigenvalues  $\lambda_1 = 4$ ,  $\lambda_2 = 1$  as well as  $\lambda_m = 0$  for  $m \geq 3$ . Note that these two eigenfunctions in fact resemble the shapes of the estimated ones in Medfly example. The predictor FPC scores are  $\xi_{im} \sim \mathcal{N}(0, \lambda_m)$ ,  $m = 1, 2$ . The measurement error  $\varepsilon_{ij}$  [(2.9) in the paper] are i.i.d.  $N(0, \sigma_x^2)$ , where two noise levels of the predictor process were considered to demonstrate the influence,  $\sigma_x = 0.1$  and  $0.3$ . Each predictor trajectory was sampled at locations that were uniformly distributed over the domain  $[0, 10]$ . The number of measurements was independently chosen for each trajectory, by selecting a number from  $\{100, \dots, 150\}$  with equal probability.

In Scenario 1 the response was generated from a single regression function,  $\beta(t) = \phi_1(t) + \phi_2(t)$  for  $t \in [0, 10]$ , with an i.i.d. additive noise  $\varepsilon_{i,y}$  distributed as  $N(0, \sigma_y^2)$  for all subjects. We also included two noise levels of the response,  $\sigma_y = 0.2$  and  $0.6$ . In Scenario 2, the response was simulated from two distinct regression functions,  $\beta_1(t) = \phi_1(t) + \phi_2(t)$  for the first 100 subjects and  $\beta_2(t) = \phi_1(t) - \phi_2(t)$  for the rest, and again was contaminated with an i.i.d. additive  $N(0, \sigma_y^2)$  noise  $\varepsilon_{i,y}$ , where  $\sigma_y = 0.2$  and  $\sigma_y = 0.6$  were considered. The proposed FMR was estimated as described in Section 2.3, including automatic choices of various smoothing parameters, the number of FPCs of the predictor processes truncated by the threshold of 90% of overall variation, and the numbers of regression functions chosen by BIC in mixture fitting. It is worth mentioning that  $M = 2$  was correctly specified in most Monte Carlo runs for each case.

We first examine model estimation using the training samples, including the regression coefficients as well as the choice of  $K$ . The benchmark we compared with is the ideal case of fitting the FMR [(2.4) in the paper] using the true FPC scores  $\xi_{im}$ . From Tables 1 and 2, one can see that, for identifying the number of regression functions  $K$ , the proposed FMR methodology is nearly as good as the ideal fitting, where the worst case is 486/500 (97.2%) in the case with the larger noise on predictor process in Scenario 2. These results also provide evidence for the consistency of regression coefficient estimates.

It is of more interest to inspect the predictive ability of the FMR when comparing with the classical functional linear models (FLM). Recall that, for each run, we have generated a validation sample of size  $n = 200$ , and here we use them to calculate the relative prediction error (RPE), defined as  $\text{RPE} = \sum_{i=1}^n (Y_i^* - \hat{Y}_i^*)^2 / \sum_{i=1}^n Y_i^{*2}$ , where  $Y_i^*$  is the response of the  $i$ th new subject in the validation sample and  $\hat{Y}_i^*$  is its predicted value. These predicted values were obtained as follows. First the FPC scores  $\hat{\xi}_{im}^*$  of the new subjects were calculated by applying the integral approach [(2.10) in the paper] to the new noisy predictor trajectories  $U_{ij}^*$ . Then these FPC scores were fed into the fitted FMRs and FLMs respectively to calculate the predicted values, where the parameters of such fitted FMRs and FLMs were estimated from the training sample. It is noticed that the response  $Y_i^*$  were used to determine which cluster the subjects belong to in FMR. From the Monte Carlo estimates of the RPEs listed in Table 3, we see that the FMR achieves dramatic gains ranging from 83% to 91% in Scenario 2, which suggests that the FMR can definitely be a viable alternative when the FLM is not adequate. For Scenario 1, comparable results were obtained for FMR and FLM. This was expected as the true value for  $K = 1$  was correctly specified by FMR in most runs. These comparisons indeed

provide strong evidence for the need of the proposed FMR when a single regression function is not sufficient to characterize the underlying relationship. Also reported in Table 3 are, for Scenario 2, the FMR predictive classification rates for the validation samples that correspond to those runs with  $K$  correctly specified as 2. As expected, they are affected by the noise levels of the predictor process and response.

## 2. THEORETICAL RESULTS

We state in this section the theoretical results on the consistency of the proposed functional mixture regression (FMR) in terms of model estimation and prediction, along with a brief and intuitive outline of the technical arguments. We first need to appropriately quantify the discrepancy between the true and estimated functional principal component (FPC) scores, i.e.,  $\xi_{im}$  and  $\hat{\xi}_{im}$ . Besides needing a large number of subjects, it is also required that the measurements sampled from each subject are sufficiently dense. Then the FPC scores can be satisfactorily estimated by the integral approximation  $\hat{\xi}_{im}^I$  [(2.10) in the paper]. Since the PACE estimates  $\hat{\xi}_{im}^P$  [(2.11) in the paper] can be considered equivalent to  $\hat{\xi}_{im}^I$  in the dense case (Müller, 2005), we shall focus on the integral estimates for theoretical developments and suppress the superscript “ $I$ ” whenever appropriate.

Write  $\boldsymbol{\xi}_i = (\xi_{i1}, \dots, \xi_{iM})^T$  and  $\hat{\boldsymbol{\xi}}_i = (\hat{\xi}_{i1}, \dots, \hat{\xi}_{iM})^T$ , where  $M$  is the number of FPCs used for approximation. We call  $\hat{X}_\xi = (\hat{\boldsymbol{\xi}}_1, \dots, \hat{\boldsymbol{\xi}}_n)^T$  the “estimated” design matrix. Given the estimated FPC scores  $\hat{\xi}_{im}$ , any estimate of the parameter  $\boldsymbol{\psi}$  [defined prior to (2.6) in the paper] would in fact be calculated from the “estimated” log-likelihood

$$l_n(\boldsymbol{\psi}; \mathbf{y}, \hat{X}_\xi) = \sum_{i=1}^n l(\boldsymbol{\psi}; y_i, \hat{\boldsymbol{\xi}}_i) = \sum_{i=1}^n \log f(y_i | \hat{\boldsymbol{\xi}}_i, \boldsymbol{\psi}) \quad (2.1)$$

instead of the “true” log-likelihood

$$l_n(\boldsymbol{\psi}; \mathbf{y}, X_\xi) = \sum_{i=1}^n l(\boldsymbol{\psi}; y_i, \boldsymbol{\xi}_i) = \sum_{i=1}^n \log f(y_i | \boldsymbol{\xi}_i, \boldsymbol{\psi}),$$

where  $f(y_i | \boldsymbol{\xi}_i, \boldsymbol{\psi})$  is defined in (2.7) in the paper. Although the consistency of the Maximum Likelihood Estimation (MLE) of  $\boldsymbol{\psi}$  obtained by maximizing the “true” likelihood is applicable to standard mixture regression (Jiang and Tanner, 1999), to the best of our knowledge, there is no existing theory for estimation obtained by maximizing the “estimated” likelihood (2.1). For clarity we denote such an estimate as  $\hat{\boldsymbol{\psi}}$  and call it MLEED, short for MLE based on the Estimated Design matrix  $\hat{X}_\xi$ . A general theorem concerning the consistency of such MLEED has been established in Yao (2010) and is stated in Lemma 3 of Section 4.

We shall consider the case of normal random component and denote the density function of a standard normal by  $\varphi(\cdot)$ . Coupling Lemmas 1–3 in Section 4, together with mild regularity conditions listed in Section 3, we have the following theorem. Recall that  $\Theta$  is the parameter space and  $f(y_i | \boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  is the  $k$ th conditional density, defined in (2.6) and (2.7) in the paper, respectively.

**Theorem 1** Suppose that the assumptions (A1)-(A4) hold with the  $k$ th conditional density  $f(y_i | \boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \varphi\{(y_i - b_{k0} - \boldsymbol{\xi}_i^T \mathbf{b}_k) / \sigma_{ky}\}$ ,  $k = 1, \dots, K$ , and that the true value  $\boldsymbol{\psi}_0$  is an interior point of the parameter space  $\Theta$ . Then, for any compact set  $E \subseteq \Theta$  containing some neighborhood of the true value  $\boldsymbol{\psi}_0$ , there exists a sequence of estimates  $\hat{\boldsymbol{\psi}} \equiv \hat{\boldsymbol{\psi}}_n$  maximizing the estimated likelihood function  $l_n(\boldsymbol{\psi}; \mathbf{y}, \hat{X}_\xi)$  on  $E$ , such that  $\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}_0$ .

Our estimates aim for the regression parameter functions  $\beta_{k,M}(t) = \sum_{m=1}^M b_{km} \phi_m(t)$

[(2.5) in the paper],  $k = 1, \dots, K$ . Let  $b_{km}^{(0)}$ ,  $\beta_{k,M}^{(0)}(t)$  and  $E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k)$  [(2.4) in the paper] be the quantities evaluated at the true values  $\boldsymbol{\psi}_0$ ,  $\phi_m$  and  $\xi_{im}$ . That is,  $\beta_{k,M}^{(0)}(t) = \sum_{m=1}^M b_{km}^{(0)}\phi_m(t)$  and  $E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k) = b_{k0}^{(0)} + \sum_{m=1}^M b_{km}^{(0)}\xi_{im}$ , where  $t \in \mathcal{T}$ ,  $k = 1, \dots, K$ ,  $i = 1, \dots, n$ . Then we can obtain consistent estimation and prediction both individually and on average.

**Theorem 2** If the assumptions in Theorem 1 hold, for any compact set  $E \subseteq \Theta$  containing some neighborhood of  $\boldsymbol{\psi}_0$ , letting  $\hat{\beta}_{k,M}(t)$  and  $\hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k)$  be the quantities evaluated at  $\hat{\phi}_m$ ,  $\hat{\xi}_{im}$  and  $\hat{\boldsymbol{\psi}}$  that maximizes  $l_n(\boldsymbol{\psi}; \mathbf{y}, \hat{X}_\xi)$  on  $E$ , i.e.,  $\hat{\beta}_{k,M}(t) = \sum_{m=1}^M \hat{b}_{km}\hat{\phi}_m(t)$  and  $\hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k) = \hat{b}_{k0} + \sum_{m=1}^M \hat{b}_{km}\hat{\xi}_{im}$ , then

$$\sup_{t \in \mathcal{T}} |\hat{\beta}_{k,M}(t) - \beta_{k,M}^{(0)}(t)| \xrightarrow{p} 0, \quad \text{for } k = 1, \dots, K, \quad (2.2)$$

$$\hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k) - E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k) \xrightarrow{p} 0, \quad \text{for } i = 1, \dots, n, \quad (2.3)$$

$$\frac{1}{n} \sum_{i=1}^n \{ \hat{E}(Y_i|X_i, M, i \in \mathcal{C}_k) - E^{(0)}(Y_i|X_i, M, i \in \mathcal{C}_k) \} \xrightarrow{p} 0. \quad (2.4)$$

*Remark.* In principle, the consistency of MLEED  $\hat{\boldsymbol{\psi}}$  as well as the predictions can be extended to FMR model with other conditional densities  $f(y_i|\boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  and/or with suitable nonlinear link functions  $g(b_{k0} + \boldsymbol{\xi}_i^T \mathbf{b}_k)$ , provided that the conditions in Lemma 3 and other necessary regularity conditions are fulfilled.

### 3. TECHNICAL ASSUMPTIONS

Necessary assumptions are listed below. Briefly, these assumptions concern the number and density of measurements per trajectory, the underlying stochastic process  $X(t)$  and the noise process  $U(t)$  that generates the observed repeated measurements  $U_{ij}$  [(2.9) in the paper], as well as various smoothing parameters and kernel functions. Let  $b = b(n)$ ,

$h = h(n)$  and  $h^* = h^*(n)$  denote the bandwidths for estimating  $\hat{\mu}$  (26),  $\hat{G}$  (27) and  $\hat{\sigma}_x$  (2) in Yao, Müller and Wang (2005).

$$(A1) \quad b \rightarrow 0, h^* \rightarrow 0, h \rightarrow 0, nb^2 \rightarrow \infty, nh^{*2} \rightarrow \infty, nh^4 \rightarrow \infty, nb^6 < \infty, \\ nh^{*6} < \infty, nh^8 < \infty, \text{ as } n \rightarrow \infty,$$

Denote the sorted time points across all subjects as  $a_0 \leq t_{(1)} \leq \dots \leq t_{(N_n)} \leq b_0$ , and  $\Delta = \max\{t_{(k)} - t_{(k-1)} : k = 1, \dots, N+1\}$ , where  $N_n = \sum_{i=1}^n n_i$ ,  $\mathcal{T} = [a_0, b_0]$ ,  $t_{(0)} = a_0$ , and  $t_{(N+1)} = b_0$ . For the  $i$ th subject, suppose that the time points  $t_{ij}$  have been ordered non-decreasingly. Let  $\Delta_i = \max\{t_{ij} - t_{i,j-1} : j = 1, \dots, n_i + 1\}$  and  $\Delta^* = \max\{\Delta_i : i = 1, \dots, n\}$ , where  $t_{i0} = a_0$  and  $t_{i,n_i+1} = b_0$ , and  $\bar{n} = n^{-1} \sum_{i=1}^n n_i$ . To obtain consistent FPC score estimates, we require both the pooled data across all subjects and the data from each subject to be dense in the time domain  $\mathcal{T}$ . For convenience, we study the asymptotics in the manner of  $\bar{n} \rightarrow \infty$  as  $n \rightarrow \infty$ , and assume that

$$(A2) \quad \Delta = O(\min\{n^{-1/2}b^{-1}, n^{-1/2}h^{*-1}, n^{-1/4}h^{-1}\}), \max\{n_i : i = 1, \dots, n\} \leq \\ C\bar{n} \text{ for some } C > 0, \text{ and } \Delta^* = O(1/\bar{n}), \text{ as } n \rightarrow \infty.$$

Denote by  $U_i(t) \stackrel{\text{i.i.d.}}{\sim} U(t)$  the distribution that generates  $U_{ij}$  for the  $i$ th subject at  $t_{ij}$ . The predictor process  $X$  and measurement  $U$  are assumed to satisfy the following conditions.

$$(A3) \quad E(\|X'\|_\infty^2) < \infty, E(\|X'^2\|_\infty^2) = o(\bar{n}_x), \sup_{t \in \mathcal{T}} E[U^4(t)] < \infty.$$

Recall that smoothing kernels  $K_1$  and  $K_2$  are compactly supported densities with zero means and finite variances. The Fourier transformations of  $K_1$  and  $K_2$  are denoted by  $\kappa_1(t) = \int e^{-iut} K_1(u) du$  and  $\kappa_2(t, s) = \int e^{-(iut+ivs)} K_2(u, v) du dv$  respectively. We require

(A4)  $\int |\kappa_1(t)|dt < \infty$ ,  $\int \int |\kappa_2(t, s)|dtds < \infty$ , i.e.,  $\kappa_1(t)$  and  $\kappa_2(t, s)$  are both absolutely integrable.

Let  $g_1(u; t)$  denote the density function of  $U(t)$ , and  $g_2(u_1, u_2; t_1, t_2)$  denote the density of  $(U(t_1), U(t_2))$ . It is assumed throughout that these density functions satisfy appropriate regularity conditions.

#### 4. AUXILIARY LEMMAS

Denote the true and estimated covariance operators by  $\mathbf{G}$  and  $\widehat{\mathbf{G}}$ , generated by  $G$  and  $\widehat{G}$  respectively; i.e.,  $\mathbf{G}(f) = \int_{\mathcal{T}} G(s, t)f(s)ds$  and  $\widehat{\mathbf{G}}(f) = \int_{\mathcal{T}} \widehat{G}(s, t)f(s)ds$  for any  $f \in L^2(\mathcal{T})$ . Define

$$\begin{aligned} D_X &= [\int_{\mathcal{T}^2} \{\widehat{G}(s, t) - G(s, t)\}^2 dsdt]^{1/2}, & \delta_m &= \min_{1 \leq j \leq m} (\lambda_j - \lambda_{j+1}), \\ M^* &= \inf\{j \geq 1 : \lambda_j - \lambda_{j+1} \leq 2D_X\} - 1, & \pi_m &= 1/\lambda_m + 1/\delta_m. \end{aligned} \quad (4.1)$$

**Lemma 1** Under (A1)-(A4) and appropriate regularity conditions for density functions  $g_1(u, t)$  and  $g_2(u_1, u_2; t_1, t_2)$ ,

$$\sup_{t \in \mathcal{T}} |\hat{\mu}(t) - \mu(t)| = O_p\left(\frac{1}{\sqrt{nb}}\right), \quad \sup_{s, t \in \mathcal{T}} |\widehat{G}(s, t) - G(s, t)| = O_p\left(\frac{1}{\sqrt{nh^2}}\right), \quad (4.2)$$

and as a consequence,  $\hat{\sigma}_x^2 - \sigma_x^2 = O_p(n^{-1/2}h^{-2} + n^{-1/2}h^{*-1})$ . Considering eigenvalues  $\lambda_m$  of multiplicity one,  $\hat{\phi}_m$  can be chosen such that,  $m = 1, \dots, M^*$ ,

$$P\left(\sup_{1 \leq m \leq M} |\hat{\lambda}_m - \lambda_m| \leq D_X\right) = 1, \quad \sup_{t \in \mathcal{T}} |\hat{\phi}_m(t) - \phi_m(t)| = O_p\left(\frac{\pi_m}{\sqrt{nh^2}}\right), \quad (4.3)$$

where  $D_X$ ,  $\pi_m$  and  $M^*$  are defined in (4.1).

The next lemma provides upper bounds for the estimation errors  $|\hat{\xi}_{im}^I - \xi_{im}|$  with some specific structure, and the derivation can be found in Müller and Yao (2008). Let  $\|f\|_\infty = \sup_{t \in \mathcal{A}} |f(t)|$  for an arbitrary function  $f$  with support  $\mathcal{A}$ ,  $\|g\| = \sqrt{\int_{\mathcal{A}} g^2(t) dt}$  for any  $g \in L^2(\mathcal{A})$ , and define

$$\begin{aligned}
\theta_{im}^{(1)} &= c_1 \|X_i\| + c_2 \|X_i X_i'\|_\infty \Delta^* + c_3, & Z_m^{(1)} &= \sup_{t \in \mathcal{T}} |\hat{\phi}_m(t) - \phi_m(t)|, \\
\theta_{im}^{(2)} &= 1 + \|\phi_m \phi_m'\|_\infty \Delta^*, & Z_m^{(2)} &= \sup_{t \in \mathcal{T}} |\hat{\mu}(t) - \mu(t)|, \\
\theta_{im}^{(3)} &= c_4 \|X_i\|_\infty + c_5 \|X_i'\|_\infty + c_6, & Z_m^{(3)} &= \|\phi_m'\|_\infty \Delta^*, \\
\theta_{im}^{(4)} &= |\sum_{j=2}^{n_i} \varepsilon_{ij} \phi_m(t_{ij})(t_{ij} - t_{i,j-1})|, & Z_m^{(4)} &\equiv 1, \\
\theta_{im}^{(5)} &= \sum_{j=2}^{n_i} |\varepsilon_{ij}|(t_{ij} - t_{i,j-1}), & Z_m^{(5)} &\equiv Z_m^{(1)},
\end{aligned} \tag{4.4}$$

for some positive constants  $c_1, \dots, c_6$  that do not depend on  $i$  or  $m$ . We note that the subscripts are mainly for notational convenience and do not necessarily reflect dependence on these indices. More importantly, we emphasize that  $\theta_{im}^{(\ell)}$  are i.i.d. over  $i$  ( $\ell = 1, 3, 4, 5$ ) or nonrandom that is free  $i$  ( $\ell = 2$ ), and that the  $Z_m^{(\ell)}$  do not depend on  $i$  for all  $\ell = 1, 2, 3, 4, 5$ .

**Lemma 2** For integral estimates of the FPC scores  $\hat{\xi}_{im}^I$  in (2.10) of the paper,

$$|\hat{\xi}_{im}^I - \xi_{im}| \leq \sum_{\ell=1}^5 \theta_{im}^{(\ell)} Z_m^{(\ell)}, \quad m = 1, \dots, M^*, \tag{4.5}$$

where  $\theta_{im}^{(\ell)}$  and  $Z_m^{(\ell)}$  are defined in (4.4), and  $M^*$  is defined in (4.1).

We aim for the consistency results for any  $M$  and  $K$ , where  $M$  and  $K$  are the numbers of FPCs and distinct regression functions in the FMR model. We state a useful theorem proved in Yao (2010) as Lemma 3 regarding the consistency of the Maximum Likelihood

Estimation based on Estimated Design (MLEED), that can be shown applicable to the proposed FMR. For convenience, we first define some conditions that are required for some relevant functions. A function  $h(\boldsymbol{\psi}, y, \boldsymbol{\xi})$  is said to satisfy the assumption (B1) at  $\boldsymbol{\psi}_1 \in \Theta$ , provided that the following holds.

(B1) There exist some functions  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  and  $c(\boldsymbol{\psi})$  such that, for all possible values of  $y, \boldsymbol{\xi}', \boldsymbol{\xi}''$ , and  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$ , where  $N_{\boldsymbol{\psi}_1}$  is some neighborhood of  $\boldsymbol{\psi}_1$ ,

$$\|h(\boldsymbol{\psi}, y, \boldsymbol{\xi}'') - h(\boldsymbol{\psi}, y, \boldsymbol{\xi}')\| \leq g(y, \boldsymbol{\xi}', \boldsymbol{\psi}) \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\| + c(\boldsymbol{\psi}) \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\|^2,$$

and  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  and  $c(\boldsymbol{\psi})$  satisfy

$$\sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} E_{(\boldsymbol{\psi}_0, \Lambda_0)} \{g^2(Y, \boldsymbol{\xi}, \boldsymbol{\psi})\} < \infty, \quad \sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} c(\boldsymbol{\psi}) < \infty,$$

where  $\boldsymbol{\psi}_0$  and  $\Lambda_0$  are the true values of  $\boldsymbol{\psi}$  and  $\Lambda$ .

A function  $q(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  is said to satisfy the set of assumptions (B2) at  $\boldsymbol{\psi}_1 \in \Theta$ , if the conditions (B2.1)–(B2.3) below hold.

(B2.1)  $q(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  is upper semicontinuous in  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$  for all  $(y, \boldsymbol{\xi})$ ;

(B2.2) There exists a function  $D(y, \boldsymbol{\xi})$  such that  $E_{(\boldsymbol{\psi}_0, \Lambda_0)} D(y, \boldsymbol{\xi}) < \infty$  and  $q(y, \boldsymbol{\xi}, \boldsymbol{\psi}) \leq D(y, \boldsymbol{\xi})$  for all  $(y, \boldsymbol{\xi})$  and  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$ ;

(B2.3) For  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$  and sufficiently small  $r > 0$ ,  $\sup_{\{\boldsymbol{\psi}': \|\boldsymbol{\psi}' - \boldsymbol{\psi}\| < r\}} q(y, \boldsymbol{\xi}, \boldsymbol{\psi}')$  is measurable in  $(y, \boldsymbol{\xi})$ .

In Lemma 3, let  $f(y|\boldsymbol{\xi}, \boldsymbol{\psi})$ ,  $\boldsymbol{\psi} \in \Theta$  denote a general conditional density function with a parameter space  $\Theta$  that is a subset of  $\mathcal{R}^p$  for some positive integer  $p$  [not restricted to the

conditional density defined in (9) of Section 2.2], and  $\hat{\xi}_i$  be any sequence of estimates of  $\xi_i$ ,  $i = 1, \dots, n$ . Denote  $l(\boldsymbol{\psi}; y, \boldsymbol{\xi}) = \log f(y|\boldsymbol{\xi}, \boldsymbol{\psi})$ .

**Lemma 3** Suppose that the true value  $\boldsymbol{\psi}_0$  is an interior point of the parameter space  $\Theta$ . Consider an arbitrary compact set  $E$  satisfying  $N_{\boldsymbol{\psi}_0} \subseteq E \subseteq \Theta$  and any  $\boldsymbol{\psi} \in E$  is an interior point of  $\Theta$ , where  $N_{\boldsymbol{\psi}_0}$  is some neighborhood of  $\boldsymbol{\psi}_0$ . Assume that

- (i) There exist some  $Z_n^{(j)}$  and  $\theta_{i,n}^{(j)}$ , where  $\theta_{i,n}^{(j)}$  are either i.i.d. realizations of positive random variables  $\theta_n^{(j)}$  or nonrandom constants with respect to  $i$ , where  $j = 1, \dots, J$ , for some finite  $J$ ,

$$\|\hat{\xi}_i - \xi_i\| \leq \sum_{j=1}^J Z_n^{(j)} \theta_{i,n}^{(j)}, \quad E\{(\theta_n^{(j)})^2\} < \infty, \quad Z_n^{(j)} \sqrt{E\{(\theta_n^{(j)})^2\}} \xrightarrow{p} 0;$$

- (ii) For any  $\boldsymbol{\psi} \in E$ ,  $l(\boldsymbol{\psi}; y, \boldsymbol{\xi})$  satisfies the assumptions (B1) at  $\boldsymbol{\psi}$  with some functions  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  and  $c(\boldsymbol{\psi})$ , where  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  satisfies the assumptions (B2) at  $\boldsymbol{\psi}$ ;
- (iii) For any  $\boldsymbol{\psi} \in E$ ,  $l(\boldsymbol{\psi}; y, \boldsymbol{\xi})$  satisfies the assumptions (B2) at  $\boldsymbol{\psi}$ ;
- (iv)  $f(y|\boldsymbol{\xi}, \boldsymbol{\psi}) = f(y|\boldsymbol{\xi}, \boldsymbol{\psi}^*)$  implies that  $\boldsymbol{\psi} = \boldsymbol{\psi}^*$  in a well-defined sense (identifiability).

Then for any sequence of the maximizer  $\hat{\boldsymbol{\psi}}$  of  $l_n(\boldsymbol{\psi}; y, \hat{X}_\xi) = \sum_{i=1}^n l(\boldsymbol{\psi}; y_i, \hat{\xi}_i)$  on the compact set  $E$ , i.e., the maximum likelihood estimates based on estimated design (MLEED), one has

$$\hat{\boldsymbol{\psi}} \xrightarrow{p} \boldsymbol{\psi}_0.$$

## 5. PROOFS OF MAIN THEOREMS

*Proof of Theorem 1.* We first verify the conditions (i)–(iv) in Lemma 3 for the FMR model with the conditional density of the  $k$ th component  $f(y_i|\boldsymbol{\xi}_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \varphi\{(y_i - b_{k0} - \boldsymbol{\xi}_i^T \mathbf{b}_k)/\sigma_{ky}\}$ ,  $k = 1, \dots, K$ , where  $\varphi(u) = \exp(-u^2/2)/\sqrt{2\pi}$  is the density function of standard normal. Given the expressions of  $Z_m^{(\ell)}$  and  $\theta_{im}^{(\ell)}$  in (4.4), one notes that  $\theta_{im}^{(\ell)}$  are i.i.d. or nonrandom w.r.t.  $i$ ,  $\ell = 1, \dots, 5$ ,  $m = 1, \dots, M$ . With (A3), it is obvious that  $E\{(\theta_{im}^{(\ell)})^2\} < \infty$  for  $\ell = 1, 2, 3$ . Due to the orthonormality of  $\phi_m$  and the independence among  $\varepsilon_{ij}$ 's, we have  $E\{(\theta_{im}^{(4)})^2\} = E\{[\sum_{j=2}^{n_i} \varepsilon_{ij} \phi_m(t_{ij})(t_{ij} - t_{i,j-1})]^2\} = \text{var}\{\sum_{j=2}^{n_i} \varepsilon_{ij} \phi_m(t_{ij})(t_{ij} - t_{i,j-1})\} = \sigma_x^2 \sum_{j=2}^{n_i} \phi_m^2(t_{ij})(t_{ij} - t_{i,j-1})^2 \leq 2\sigma^2 \Delta^* \rightarrow 0$ . For  $\theta_{im}^{(5)}$ , applying Cauchy-Schwartz inequality,  $E\{(\theta_{im}^{(4)})^2\} \leq \{\sum_{j=2}^{n_i} E(\varepsilon_{ij}^2)(t_{ij} - t_{i,j-1})\} \mathcal{T} \leq 2\mathcal{T}^2 \sigma_x^2 < \infty$  for large  $n$ . Combining with Lemmas 1 and 2, then condition (i) holds. Since the parameter space  $\Theta$  defined is an open subset of  $\mathcal{R}^{(M+3)K-1}$ , any  $\boldsymbol{\psi} \in E$  is always an interior point of  $\Theta$ . It is easy to verify that condition (iii) holds for the conditional density  $f(y_i|\boldsymbol{\xi}_i, \boldsymbol{\psi})$  with normal components, while condition (iv) is satisfied given the identifiability in the sense of (2.8) in the paper.

Now we check condition (ii), and observe that

$$l(\boldsymbol{\psi}; y, \boldsymbol{\xi}) = \log \left\{ \sum_{k=1}^K \pi_k f(y|\boldsymbol{\xi}, b_{k0}, \mathbf{b}_k, \sigma_{ky}) \right\},$$

$$f(y|\boldsymbol{\xi}, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) = \frac{1}{\sqrt{2\pi}\sigma_{ky}} \exp \left\{ -\frac{(y - b_{k0} - \boldsymbol{\xi}^T \mathbf{b}_k)^2}{2\sigma_{ky}^2} \right\}. \quad (5.1)$$

For any fixed interior point  $\boldsymbol{\psi}_1$  of  $\Theta$ , one can always assume that a sufficiently small neighborhood  $N_{\boldsymbol{\psi}_1}$  is bounded, and particularly  $\delta \leq \pi_k \leq 1 - \delta$  and  $\sigma_{ky} > \delta$  for some

$\delta > 0, k = 1, \dots, K$ . By Mean Value Theorem, one has, for  $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$ ,

$$\begin{aligned} l(\boldsymbol{\psi}; y, \boldsymbol{\xi}'') - l(\boldsymbol{\psi}; y, \boldsymbol{\xi}') &= (\partial^T l(\boldsymbol{\psi}; y, \boldsymbol{\xi}^*) / \partial \boldsymbol{\xi})(\boldsymbol{\xi}'' - \boldsymbol{\xi}'), \\ \frac{\partial}{\partial \boldsymbol{\xi}} l(\boldsymbol{\psi}; y, \boldsymbol{\xi}^*) &= \frac{\sum_{k=1}^K \pi_k f(y | \boldsymbol{\xi}^*, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2) (y - b_{k0} - \boldsymbol{\xi}^{*T} \mathbf{b}_k) \mathbf{b}_k / \sigma_{ky}^2}{\sum_{k=1}^K \pi_k f(y | \boldsymbol{\xi}^*, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)}, \end{aligned}$$

where  $\boldsymbol{\xi}^* = \boldsymbol{\xi}' + v(\boldsymbol{\xi}'' - \boldsymbol{\xi}')$  for some  $0 \leq v \leq 1$ . In spite of the complex appearance of the above expression, one can see that it is in fact a *weighted average* of  $(y - b_{k0} - \boldsymbol{\xi}^{*T} \mathbf{b}_k) \mathbf{b}_k / \sigma_{ky}^2$  with weights  $\pi_k f(y | \boldsymbol{\xi}^*, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$ ,  $k = 1, \dots, K$ . Therefore,

$$\begin{aligned} \left\| \frac{\partial}{\partial \boldsymbol{\xi}} l(\boldsymbol{\psi}; y, \boldsymbol{\xi}^*) \right\| &\leq \sum_{k=1}^K \left\| (y - b_{k0} - \boldsymbol{\xi}^{*T} \mathbf{b}_k) \mathbf{b}_k / \sigma_{ky}^2 \right\| \\ &\leq \sum_{k=1}^K \frac{1}{\sigma_{ky}^2} \left\{ \|\mathbf{b}_k y - b_{k0} \mathbf{b}_k\| + \|\mathbf{b}_k\|^2 \|\boldsymbol{\xi}' + v(\boldsymbol{\xi}'' - \boldsymbol{\xi}')\| \right\} \\ &\leq \sum_{k=1}^K \frac{\|\mathbf{b}_k\|}{\sigma_{ky}^2} \left\{ |y - b_{k0}| + \|\mathbf{b}_k\| \|\boldsymbol{\xi}'\| \right\} + \left\{ \sum_{k=1}^K \frac{\|\mathbf{b}_k\|^2}{\sigma_{ky}^2} \right\} \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\| \\ &\equiv g(y, \boldsymbol{\xi}', \boldsymbol{\psi}) + c(\boldsymbol{\psi}) \|\boldsymbol{\xi}'' - \boldsymbol{\xi}'\|. \end{aligned}$$

From the boundedness of the small  $N_{\boldsymbol{\psi}_1}$ , it is easy to see that  $\sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} c(\boldsymbol{\psi}) < \infty$ ,  $\sup_{\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}} E_{(\boldsymbol{\psi}_0, \Lambda_0)} \{g^2(Y, \boldsymbol{\xi}, \boldsymbol{\psi})\} < \infty$ , and moreover  $g(y, \boldsymbol{\xi}, \boldsymbol{\psi})$  satisfies the assumptions (B2) at  $\boldsymbol{\psi}_1$ . Thus condition (ii) holds. The existence of a consistent sequence  $\hat{\boldsymbol{\psi}} \in E$  that are roots of  $\partial l_n(\boldsymbol{\psi}; y, \hat{\boldsymbol{\xi}}) / \partial \boldsymbol{\psi} = 0$  follows for the conditional mixture normal density (5.1).

*Proof of Theorem 2.* The uniform consistency of  $\hat{\beta}_{k,M}(t)$  w.r.t.  $t \in \mathcal{T}$  is obvious given Theorem 1 and Lemma 1. For individual prediction, note that  $|\hat{E}(Y_i | X_i, M) - E(Y_i | X_i, M)| \leq |b_{k0} - b_{k0}| + \|\mathbf{b}_k - \mathbf{b}_k\| \cdot \|\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\|$  and  $\|\hat{\boldsymbol{\xi}}_i - \boldsymbol{\xi}_i\| \leq \sum_{m=1}^M \sum_{\ell=1}^5 Z_m^{(\ell)} \theta_{im}^{(\ell)}$ .

We have shown that  $E\{(\theta_m^{(\ell)})^2\} < \infty$  and  $Z_m^{(\ell)} \sqrt{E\{(\theta_m^{(\ell)})^2\}} \xrightarrow{p} 0$ , where  $\theta_{im}^{(\ell)} \stackrel{\text{i.i.d.}}{\sim} \theta_m^{(\ell)}$  (considering i.i.d. random variables here without loss of generality),  $\ell = 1, \dots, 5$ ,  $m = 1, \dots, M$ . We then arrive at the result (2.3) by observing the following for each  $m$  and  $\ell$ . For any  $\epsilon > 0$  and  $\delta > 0$ , we choose  $A \geq \sqrt{2/(\epsilon\delta)}$ , i.e.,  $(A^2\delta)^{-1} \leq \epsilon/2$ , when  $n$  is sufficiently large, and apply Chebyshev's inequality:

$$\begin{aligned} P\left(Z_m^{(\ell)}|\theta_{im}^{(\ell)} - E\theta_m^{(\ell)}| > \delta\right) &\leq P\left(Z_m^{(\ell)}\sqrt{E\{(\theta_m^{(\ell)})^2\}} > \frac{\sqrt{\delta}}{A}\right) + P\left(\frac{|\theta_{im}^{(\ell)} - E\theta_m^{(\ell)}|}{A\sqrt{E\{(\theta_m^{(\ell)})^2\}}} > \sqrt{\delta}\right) \\ &\leq \frac{\epsilon}{2} + \frac{1}{A^2\delta} \leq \epsilon. \end{aligned}$$

Noting that  $(1/n)\|\hat{\xi}_i - \xi_i\| \leq \sum_{m=1}^M \sum_{\ell=1}^5 Z_m^{(\ell)}(1/n) \sum_{i=1}^n \theta_{im}^{(\ell)}$ , then the consistency of the average prediction (2.4) follows immediately from the law of large numbers for triangular arrays.

## 6. EM ALGORITHM FOR MIXTURE REGRESSION MODELS

For completeness we outline an EM algorithm for fitting mixture regression models. For details, see, for example, Naik, Shi and Tsai (2007).

Consider the following mixture model with  $K$  normal density components:

$$f(y_i|\xi_i, \psi) = \sum_{k=1}^K \pi_k \varphi(y_i|\xi_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2), \quad i = 1, \dots, n,$$

where  $0 < \pi_k < 1$  and  $\sum \pi_k = 1$ ,  $\varphi(y_i|\xi_i, b_{k0}, \mathbf{b}_k, \sigma_{ky}^2)$  is the normal density with mean  $(b_{k0} + \xi_i^T \mathbf{b}_k)$  and variance  $\sigma_{ky}^2$ , and  $\psi = (\mathbf{b}_0^T, \mathbf{b}_1^T, \dots, \mathbf{b}_K^T, \boldsymbol{\pi}^T, \boldsymbol{\sigma}_y^T)^T$  is the vector containing all relevant parameters. Let  $\psi^{(r)}$  be the  $r$ -th iterative estimate for  $\psi$ . In the

E-Step of the algorithm, one calculates

$$\tau_{ik}^{(r)} = \frac{\pi_k^{(r)} \varphi(y_i | \boldsymbol{\xi}_i, b_{k0}^{(r)}, \mathbf{b}_k^{(r)}, \sigma_{ky}^{2(r)})}{\sum_{k=1}^K \pi_k^{(r)} \varphi(y_i | \boldsymbol{\xi}_i, b_{k0}^{(r)}, \mathbf{b}_k^{(r)}, \sigma_{ky}^{2(r)})}.$$

This quantity can be seen as the  $r$ -th estimated probability for  $y_i$  originated from the  $k$ -th component.

In the M-Step, the  $(r+1)$ -th estimates are calculated with the following closed-form expressions:  $\pi_k^{(r+1)} = n^{-1} \sum_{i=1}^n \tau_{ik}^{(r)}$  and

$$\begin{pmatrix} b_{k0}^{(r+1)} \\ \mathbf{b}_k^{(r+1)} \end{pmatrix} = (\tilde{X}_k^{(r)T} \tilde{X}_k^{(r)})^{-1} \tilde{X}_k^{(r)T} \tilde{\mathbf{y}}_k^{(r)}, \quad \sigma_{ky}^{2(r+1)} = \frac{\tilde{\mathbf{y}}_k^{(r)T} (I - \tilde{H}_k^{(r)}) \tilde{\mathbf{y}}_k^{(r)}}{\text{tr}(W_k^{(r)})},$$

for  $k = 1, \dots, K$ . In the above  $W_k^{(r)} = \text{diag}(\tau_{1k}^{(r)}, \dots, \tau_{nk}^{(r)})$ ,  $\tilde{X}_k^{(r)} = W_k^{(r)1/2} \tilde{X}_\xi$ ,  $\tilde{\mathbf{y}}_k^{(r)} = W_k^{(r)1/2} \mathbf{y}$ ,  $\tilde{H}_k^{(r)} = \tilde{X}_k^{(r)} (\tilde{X}_k^{(r)T} \tilde{X}_k^{(r)})^{-1} \tilde{X}_k^{(r)T}$ ,  $\tilde{X}_\xi = (\mathbf{1}, X_\xi)$ ,  $X_\xi = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n)^T$  and  $\mathbf{y} = (y_1, \dots, y_n)^T$ .

## REFERENCES

- JIANG, W. AND TANNER, M. A. (1999). Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Annals of Statistics* **27**, 987–1011.
- MÜLLER, H. G. (2005). Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32**, 223–240.
- MÜLLER, H. G. AND YAO, F. (2008). Functional additive models. *Journal of the American Statistical Association* **103**, 1534–1544.
- NAIK, P. A., SHI, P. AND TSAI, C. L. (2007). Extending the Akaike Information Criterion to Mixture Regression Models. *Journal of the American Statistical Association* **102**, 244–254.
- YAO, F. (2010). Maximum likelihood estimation based on estimated design or data. *Technical Report*.

YAO, F., MÜLLER, H. G. AND WANG, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 577–590.

Table 1. *Section 3 Simulation Scenario 1. Monte Carlo estimates of regression coefficients (standard errors in parentheses) for 4 combinations of noise levels, calculated from those runs that  $K = 1$  were correctly specified. The first integer in each case reports the number, out of 500 runs, of correctly specified runs. In this scenario there is one single regression function with true values  $b_{11} = b_{12} = 1$ . The first row corresponds to the ideal fitting (IDEAL) while the second row corresponds to FMR.*

		Noises	$\sigma_x = .1$		$\sigma_x = .3$	
IDEAL	$\sigma_y = .2$	498	.9996	.9998	496	.9998 1.0002
FMR		495	.9796	.9975	494	.9179 .9982
			(.0073)	(.0152)		(.0049) (.0070)
			(.0493)	(.0546)		(.0798) (.0770)
IDEAL	$\sigma_y = .6$	497	1.0002	1.0004	496	1.0002 1.0009
FMR		497	.9908	.9934	497	.9251 .9942
			(.0141)	(.0206)		(.0147) (.0216)
			(.0861)	(.0867)		(.0835) (.0854)

Table 2. *Similar to Table 1 but for Scenario 2. For this scenario the true value for  $K = 2$  and the regression coefficients are  $(b_{11}, b_{12}, b_{21}, b_{22}) = (1, 1, 1, -1)$ .*

		Noises	$\sigma_x = .1$				$\sigma_x = .3$				
IDEAL	$\sigma_y = .2$	494	.9998	.9999	.9999	-1.0007	495	.9998	1.0004	.9997	-.9999
FMR		494	.9786	.9992	.9780	-1.0014	486	.9969	.8422	1.0114	-.8211
			(.0111)	(.0217)	(.0110)	(.0208)		(.0067)	(.0100)	(.0072)	(.0101)
			(.0519)	(.0628)	(.0531)	(.0603)		(.0832)	(.0838)	(.0827)	(.0839)
IDEAL	$\sigma_y = .6$	493	1.0004	1.0008	1.0009	-1.0009	496	1.0009	.9988	1.0001	-.9980
FMR		492	.9922	.9937	.9842	-1.0013	488	.9261	.9933	.9197	-.9979
			(.02189)	(.0293)	(.0235)	(.0315)		(.0222)	(.0317)	(.0225)	(.0310)
			(.0895)	(.0880)	(.0887)	(.0922)		(.0913)	(.0954)	(.0945)	(.0953)

Table 3. Monte Carlo estimates of the relative prediction errors (RPE) defined in Section 3 for 4 combinations of noise levels. Also reported in the last row is the predictive classification rates (P. C. Rate) calculated for the validation samples that correspond to those runs with  $K = 2$  correctly specified.

Model	Method	Noise levels: $\{\sigma_x, \sigma_y\}$			
		$\{.1, .2\}$	$\{.1, .6\}$	$\{.3, .2\}$	$\{.3, .6\}$
Scenario I ( $K = 1$ )	FLM	.02448	.03408	.05764	.08152
	FMR	.02447	.03408	.05764	.08152
Scenario II ( $K = 2$ )	FLM	.23210	.34332	.37007	.39004
	FMR	.0218	.03016	.04943	.06804
	(P. C. Rate)	(.8932)	(.8951)	(.8664)	(.8448)