

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 ε_{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 ξ_{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., ξ_{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}_ξ . 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 Θ 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 Θ . 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 $\boldsymbol{\psi}_0$, ϕ_m and ξ_{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 λ_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 , π_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 Λ_0 are the true values of $\boldsymbol{\psi}$ and Λ .

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 Θ 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 ξ_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 Θ . 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 Θ , 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 ϕ_m and the independence among ε_{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 Θ defined is an open subset of $\mathcal{R}^{(M+3)K-1}$, any $\boldsymbol{\psi} \in E$ is always an interior point of Θ . 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 Θ , 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 $\boldsymbol{\psi} \in N_{\boldsymbol{\psi}_1}$,

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

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 ℓ . 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 σ_{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 ψ . 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	FLM	.02448	.03408	.05764	.08152
($K = 1$)	FMR	.02447	.03408	.05764	.08152
Scenario II	FLM	.23210	.34332	.37007	.39004
($K = 2$)	FMR	.0218	.03016	.04943	.06804
	(P. C. Rate)	(.8932)	(.8951)	(.8664)	(.8448)