Statistica Sinica **33** (2023), 2087-2115
doi:https://doi.org/10.5705/ss.202021.0183

FUNCTIONAL FINITE MIXTURE REGRESSION MODELS

Xiao Wang[1], Leo Yu-Feng Liu[2] and Hongtu Zhu[2]

[1]*Purdue University and* [2]*University of North Carolina at Chapel Hill*

Abstract: The aim of this study is to develop a set of functional finite mixture regression models with functional predictors in the framework of the reproducing kernel Hilbert space. First, we show the consistency of a penalized likelihood model order estimator for the true model order, denoted as q^*. We further show that the penalty of order $q^{2r/(2r+1)}n^{1/(2r+1)}\log(n)$ yields a strong consistent estimator of q^*, where n and q are the sample size and the model order, respectively, and r is the eigenvalue decay rate of an operator determined jointly by the reproducing and covariance kernels. Second, we establish the minimax rate of convergence for the estimation risk. We show that the optimal rate is determined by the alignment of the reproducing kernel and the covariance kernel and the true model order q^*. An efficient algorithm is also developed to estimate all unknown components of the functional finite mixture model. Simulation studies and a real-data analysis illustrate the merits of the proposed method.

1. Introduction

With the rapid growth of technology, many large-scale biomedical studies (e.g., UK Biobank) have collected massive data sets with large volumes of multimodality imaging, genetic, neurocognitive, and clinical information from increasingly large cohorts. Simultaneously extracting and integrating rich and diverse heterogeneous information in neuroimaging and/or other variables from these big data sets could transform our understanding of how diseases impact the structure and function of the brain and cognitive functions across the human lifespan. Therefore, it is imperative to develop new learning methods (Liu and Zhu (2021); Feng et al. (2020); Wang et al. (2021); Ombao et al. (2016)) that are applicable to neuroimaging studies for neuropsychiatric disorders, major neurodegenerative diseases, and normal brain development.

Mixture models are powerful probabilistic models that use mixture distributions to represent the presence of subpopulations within an overall population; for a comprehensive review of the theory and applications of mixture models, see McLachlan and Peel (2000). The most well-known mixture model is the Gaussian

Corresponding author: Hongtu Zhu, Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7400, USA. E-mail: htzhu@email.unc.edu.

mixture model (Richard and Green (1997)), in which the density of a random vector Y is represented by $p(y) = \sum_{i=1}^{q} \phi_i N(\mu_i, \Sigma_i)$, where the ith vector component is characterized by a normal distribution with weight ϕ_i, mean μ_i, and covariance Σ_i. This framework can be extended naturally to a regression setting. The Gaussian mixture regression model (Ghahramani and Jordan (1994); Calinon, F. and Billard (2007); Stulp and Sigaud (2015)) assumes that the conditional distribution of multiple responses given one or more predictors follows a Gaussian mixture distribution, and that the conditional mean is a function of the predictors. In this study, we investigate major mathematical challenges in the conditional analysis of clinical response variables given ultrahigh-dimensional imaging predictors under the framework of functional finite mixture regression models.

Functional data analysis has been an active area of research, and well-known monographs in this area include those of Ramsay and Sloverman (2005), Bowman (2010), and Ferraty and Vieu (2006). Functional regression, and particularly the functional linear regression model, has been studied extensively. For example, the scalar-on-function regression (a continuous response variable regressed on functional covariates), which corresponds to our proposed mixture model (1.1) with a single component, has been studied by Cai and Hall (2006), Crambes, Kneip and Sarda (2009), Wang, Chiou and Muller (2016), Yuan and Cai (2010), Hall and Horowitz (2007), Du and Wang (2014) and Wang, Zhu and ADNI (2017). When the true model is a finite mixture model with at least two different components, the standard functional linear regression provides poor results. Few studies have examined numerical algorithms for functional finite mixture regression models. Yao, Fu and Lee (2011) proposed a class of functional regression models that allow the regression structure to vary for different subpopulations. By projecting the functional predictor process onto its eigenspace, the new functional finite mixture regression model is simplified to a framework that is similar to classical mixture regression models. However, no studies have considered the reproducing kernel Hilbert space (RKHS) or investigated the theoretical properties of RKHS estimates, particularly the minimax rate.

Consider a functional finite mixture regression model such that the conditional distribution of a scalar response Y given a functional predictor $\{X(t) : t \in \mathcal{I}\}$ belongs to the class of convex combinations of q^* densities given by

$$\mathcal{M}_{q^*} = \left\{ f^*(y|x) = \sum_{k=1}^{q^*} \pi_k^* f_0\Big(y - \langle x, \beta_k^* \rangle\Big) : \right.$$

$$\pi_k^* \geq 0, \sum_{k=1}^{q^*} \pi_k^* = 1, \beta_k^* \in \mathcal{F} \Bigg\}, \tag{1.1}$$

where f_0 is a fixed probability density (e.g., Gaussian) on $\mathbb{R}$ and $\langle x, \beta \rangle = \int_{\mathcal{I}} x(t)\beta(t)dt$. The intercept terms can be incorporated into the model by writing each component as $f_0(y - \beta_{0k}^* - \langle x, \beta_k^* \rangle)$, where β_{0k}^* are intercepts. We ignore them in the following analysis only for ease of presentation. Here, $\mathcal{F}$ is assumed to be an RKHS with the reproducing kernel K. The RKHS $\mathcal{F}$ is a linear functional space endowed with an inner product $\langle \cdot, \cdot \rangle_K$ such that for any $t \in \mathcal{I}$, $K(t, \cdot) \in \mathcal{F}$, and $f(t) = \langle K(t, \cdot), f \rangle_K$ holds for any $f \in \mathcal{F}$. For more details on the RKHS, please see Wahba (1990), Steinwart and Christmann (2008), Schölkopf (2001) and the references therein. Here, $\mathcal{M}_{q^*}$ is parametrized by the parameters $\Pi_{q^*}^* = (\pi_1^*, \ldots, \pi_{q^*}^*) \in \Delta_{q^*-1}$ and $\Theta_{q^*}^* = (\beta_1^*, \ldots, \beta_{q^*}^*) \in \mathcal{F}^{q^*}$, where Δ_{q^*-1} is the q^*-simplex. We call q^* the *number of mixture components* or the *model order*, $\Pi_{q^*}^*$ the *mixing probabilities*, and $\Theta_{q^*}^*$ the *coefficient functions*.

Suppose that we observe n independent and identically distributed (i.i.d.) copies of (Y, X), denoted as $\{(Y_i, X_i) : i = 1, \ldots, n\}$. Let $\mathcal{M} = \cup_q \mathcal{M}_q$. Let $f^*(y|x) \in \mathcal{M}_{q^*}$ be the true conditional distribution and $\widehat{f}(y|x) \in \mathcal{M}$ be the estimated conditional density. We are interested in investigating three important questions for model (1.1):

(a) how to construct an estimator $\widehat{f}(y|x)$ that can achieve the optimal minimax rate of convergence;

(b) how to consistently estimate the number of mixture components q^*; and

(c) how to numerically compute all unknown parameters and functions in $\widehat{f}(y|x)$.

These questions are known to present major challenges, even for parametric mixture regression models (Chernoff (1954); Dacunha-Castelle and Gassiat (1999); Zhu and Zhang (2004); Ho and Nguyen (2016); Heinrich and Kahn (2018)). The accuracy of an estimation can be measured naturally using

$$\mathcal{R}_n(\widehat{f}, f^*) = \mathbb{E}_X \Bigg\{ \int \left(\sqrt{\widehat{f}(y|X)} - \sqrt{f^*(y|X)} \right)^2 dy \Bigg\}, \tag{1.2}$$

which is the squared Hellinger distance between $\widehat{h}(y, x) = \widehat{f}(y|x) f_X(x)$ and $h^*(y, x) = f^*(y|x) f_X(x)$. As the sample size n increases, the convergence rate of $\mathcal{R}_n$ reflects the difficulty of the estimation problem.

We carry out a systematic investigation of model (1.1). We make four major contributions to the literature:

(i) We construct an optimal estimate of $f^*(y|x)$ using the following steps. We show that a minimax lower bound of $\mathcal{R}_n$ in (1.2) depends on the reproducing kernel K, covariance function C of the random predictor X, and model order q^*. In particular, it depends on the decay order of the eigenvalues of the operator $K^{1/2}CK^{1/2}$. A similar phenomenon has been found in works on functional regressions (Cai and Yuan (2012); Du and Wang (2014); Wang and Ruppert (2015)). Then, we establish a minimax upper bound of $\mathcal{R}_n$ based on a penalized likelihood estimator when the true coefficient functions reside in $\mathcal{M}_q$, with $q \geq q^*$. We also propose a minimal penalty that yields a consistent order estimation.

(ii) We propose a general class of penalized likelihood order estimators in order to select and estimate the model order q^*. Theoretically, we establish the strong consistency of the order estimators for model (1.1). In contrast, most existing consistency results assume a prior upper bound on the order (Csiszár and Shields (2000); Nishii (1988)). The only exception that we are aware of is the work of (Gassiat and van Handel (2012)), which explores the consistency properties of the penalized likelihood model order estimator and provides the minimal strong consistency penalty.

(iii) We develop an expectation-maximization (EM) algorithm to estimate all unknown parameters and functions.

(iv) We examine the finite-sample performance of our methods by using simulations and a real-data set collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study.

The remainder of this paper is structured as follows. Section 2 establishes the minimax rate of convergence of $\mathcal{R}_n(\widehat{f}, f^*)$ in (1.2). Section 3 presents the estimation of q^* and the estimation of the coefficient functions $\Theta^*_{q^*}$. Section 4 summarizes the results of our numerical experiments and real-data analysis. Section 5 concludes the paper.

2. Methodology

In this section, we first establish the minimax rate of convergence of $\mathcal{R}_n(\widehat{f}, f^*)$ in (1.2), and then introduce an estimation algorithm for $\widehat{f}$.

2.1. Optimal rate of convergence

The optimal rate of convergence of $\mathcal{R}_n(\widehat{f}, f^*)$ is established in several steps. We first derive a minimax lower bound, and then show that the convergence rate

of the lower bound is optimal by constructing an estimator that can attain this rate of convergence.

Minimax lower bound

We first establish the minimax lower bound of $\mathcal{R}_n$, based on the following assumption.

A1. Let $C(t,s) = \mathrm{cov}(X(t), X(s))$ be the covariance function of X, and $\{\rho_k : k \in \mathbb{N}\}$ be the nonincreasing ordered eigenvalues of the operator $K^{1/2}CK^{1/2}$. Assume that $\rho_k \asymp k^{-2r}$, with $r > 0$.

Assumption A1 specifies that the decay rate of the eigenvalues of the operator $K^{1/2}CK^{1/2}$ is of order k^{-2r}. As a concrete example, consider the univariate Sobolev space W_2^m. It is known that its reproducing kernel is $(m!)^{-2}B_m(s)B_m(t)+(-1)^{m-1}\{(2m)!\}^{-1}B_{2m}(|s-t|)$, where B_m is the mth Bernoulli polynomial (Wahba (1990)). It is also known that the decay rate of the eigenvalues of this kernel is of order k^{-2m} (Micchelli and Wahba (1981)).

The minimax lower bound is given in the following theorem.

Theorem 1. *Assuming* A1 *holds, we have*

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\widehat{f}}\sup_{f^*\in\mathcal{M}_{q^*}}\mathbb{P}\left(\mathcal{R}_n(\widehat{f}, f^*) \geq a\left(\frac{q^*}{n}\right)^{2r/(2r+1)}\right) = 1. \tag{2.1}$$

The result in Theorem 1 is an asymptotic result. It shows that there exists a function $V_1(a)$ such that $\lim_{n\to\infty}\inf_{\widehat{f}}\sup_{f^*}\mathbb{P}(\mathcal{R}_n(\widehat{f}, f^*) \geq a(q^*/n)^{2r/(2r+1)}) \geq V_1(a)$ and $V_1(a) \to 1$ as $a \to 0$. The proof of Theorem 1 is provided in the appendix. The main tool is adopted from Tsybakov (2009) by realizing that any lower bound for a specific case immediately yields a lower bound for the general case. Theorem 1 shows that the minimax lower bound depends on the model order q^* and how the reproducing kernel K and the covariance function C are aligned. In general, the eigenvalues of K and C alone cannot determine the decay rate of the eigenvalues of $K^{1/2}CK^{1/2}$. For functional linear regression models, Cai and Yuan (2012) used the same tool to establish the minimax lower bound for prediction.

Minimax upper bound with $q \geq q^*$

Let $f^*(y|x) \in \mathcal{M}_{q^*}$ be the true conditional distribution and $\mathcal{M} = \cup_{q=1}^{\infty}\mathcal{M}_q$. The log-likelihood function can be written as

$$\ell_n(f) = \sum_{i=1}^{n}\log f(Y_i|X_i)$$

$$= \sum_{i=1}^{n} \log \sum_{k=1}^{q} \pi_k f_0(Y_i - \eta(X_i, \beta_k)), \qquad f \in \mathcal{M}, \tag{2.2}$$

where $\eta(X, \beta) = \int_{\mathcal{I}} X(t)\beta(t)dt$. We consider a general class of penalized likelihood estimations given by

$$\widehat{f} = \underset{f \in \mathcal{M}_q}{\operatorname{argmin}} \Big\{ -\ell_n(f) + \lambda \|\beta\|_K^2 \Big\}, \tag{2.3}$$

for $q \geq q^*$, where λ is a smoothing parameter that balances the trade-off between the goodness of fit to the data and the smoothness of the estimator, and $\|\cdot\|_K$ is the RKHS norm.

The following notation is used throughout:

$$H_0(y|x) = \sup_{\eta \in \mathbb{R}} \frac{f_0(y-\eta)}{f^*(y|x)}, \quad H_1(y|x) = \sup_{\eta \in \mathbb{R}} \frac{|\dot{f}_0(y-\eta)|}{f^*(y|x)},$$
$$H_2(y|x) = \sup_{\eta \in \mathbb{R}} \frac{|\ddot{f}_0(y-\eta)|}{f^*(y|x)}, \quad H_3(y|x) = \sup_{\eta \in \mathbb{R}} \frac{|f_0^{(3)}(y-\eta)|}{f^*(y|x)},$$

where $\dot{f}_0$, $\ddot{f}_0$, and $f_0^{(3)}$ are the first-, second-, and third-order derivatives of $f_0(\cdot)$, respectively. We need two additional assumptions.

B1. $f_0 \in C^3$ and $f_0(x)$ and $\dot{f}_0(x)$ vanish as $x \to \infty$. We assume $H_k(\cdot|\cdot) \in L^4(h^* d\mu)$ for $k = 0, 1, 2$ and $H_3(\cdot|\cdot) \in L^2(h^* d\mu)$.

B2. Given the functional predictor X, there exists a disjoint partition of $\mathbb{R} = A_0 \cup \{\cup_{k=1}^{q^*} A_k\}$ such that $A_1, \ldots, A_{q^*}$ are bounded intervals, where each bounded interval A_j contains precisely one component $\eta_k^* = \langle X, \beta_k^* \rangle$, for $k = 1, \ldots, q^*$, and the unbounded set A_0 contains no component.

Condition B1 is similar to the condition adopted in Gassiat and van Handel (2014). This condition is satisfied in particular when f_0 is chosen as the standard normal density. Condition B2 characterizes the identifiability issue. In general, given x, the mean components $\langle x, \beta_k^* \rangle$ should be well separated to guarantee identifiability.

Theorem 2. *Assume that* B1–B2 *hold. Let* $\{\rho_k : k \in \mathbb{N}\}$ *be the nonincreasing ordered eigenvalues of the operator* $K^{1/2}CK^{1/2}$. *Assume that* $\rho_k \asymp k^{-2r}$, *with* $r > 1/2$. *For any* $q \geq q^*$, *we have*

$$\lim_{A \to \infty} \lim_{n \to \infty} \sup_{f^* \in \mathcal{M}_{q^*}} \mathbb{P}\left(\mathcal{R}_n(\widehat{f}, f^*) \leq A \left(\frac{q}{n} \right)^{2r/(2r+1)} \right) = 1 \tag{2.4}$$

and $\|\widehat{\beta}\|_K^2 = O_p(1)\|\beta^*\|_K^2$*, provided that* λ *is of order* $q^{2r/(2r+1)}n^{1/(2r+1)}$.

The result in Theorem 2 is also an asymptotic result. It shows that there exists a function $V_2(a)$ such that $\lim_{n\to\infty} \inf_{\widehat{f}} \sup_{f^*} \mathbb{P}(\mathcal{R}_n(\widehat{f}, f^*) \leq A(q^*/n)^{2r/(2r+1)}) \geq V_2(A)$ and $V_2(A) \to 1$ hold as $A \to \infty$. The proof of Theorem 2 is provided in the appendix. It combines two main tools. The first one characterizes the local geometry of finite mixtures (Gassiat and van Handel (2014)), and the second uses the empirical process theory with covering numbers to establish the convergence rate (van de Geer (2000)). The penalized estimator has an explicit solution for the function linear regression models in Cai and Yuan (2012), and establishing the upper bound is more straightforward. Theorem 2 establishes the upper bound on the rate of convergence of $\mathcal{R}_n$ based on the penalized likelihood estimator when $q \geq q^*$. However, we still cannot claim the minimax optimal rate of convergence, because the model order q^* is unknown. Therefore, a consistent estimation of q^* is critical to achieving the minimax optimality on the rate of convergence of $\mathcal{R}_n$.

Consistent order estimation

The number of components q^* of the true mixture $f^* \in \mathcal{M}$ can be estimated by using a general class of penalized likelihood order estimators:

$$\widehat{q}_n = \operatorname*{argmax}_{q\geq 1} \left\{ \sup_{f\in\mathcal{M}_q} \left(\ell_n(f) - \lambda\|\beta\|_K^2\right) - \operatorname{pen}_n(q) \right\}, \tag{2.5}$$

where $\operatorname{pen}_n(q)$ is a penalty function and $\ell_n(f)$ is the likelihood function. Our goal is to understand which $\operatorname{pen}_n(q)$ yields the strong consistency of the order estimator, that is, $\widehat{q}_n \to q^*$ as $n \to \infty$ a.s. Achieving this goal requires a precise understanding of the difference of the penalized log-likelihood functions given by

$$\sup_{f\in\mathcal{M}_q} \left(\ell_n(f) - \lambda_{q,n}\|\beta\|_K^2\right) - \sup_{f\in\mathcal{M}_{q^*}} \left(\ell_n(f) - \lambda_{q^*,n}\|\beta\|_K^2\right)$$

as $n \to \infty$, uniformly in the model order $q > q^*$. In this section, we discuss the selection of $\operatorname{pen}_n(q)$ that yields the strong consistency of $\widehat{q}_n$.

Theorem 3. *Assume* B1–B2 *hold. Let* $\{\rho_k : k \in \mathbb{N}\}$ *be the nonincreasing ordered eigenvalues of the operator* $K^{1/2}CK^{1/2}$*. Assume that* $\rho_k \asymp k^{-2r}$*, with* $r > 1/2$*. Let* $c > 0$ *be a constant and*

$$pen_n(q) = cq^{2r/(2r+1)}n^{1/(2r+1)}\log(n). \tag{2.6}$$

Then, we have

(a) $\widehat{q}_n \to q^*$ *as* $n \to \infty$ *a.s.;*

(b) $\lim_{A\to\infty}\lim_{n\to\infty}\sup_{f^*\in\mathcal{M}_{q^*}}\mathbb{P}(\mathcal{R}_n(\widehat{f},f^*)\geq A(q^*/n)^{2r/(2r+1)})=0.$

The proof of Theorem 3 is provided in the appendix, and is obtained by precisely characterizing the difference of the penalized log-likelihood functions. Intuitively, because we are dealing with infinite-dimensional parameters such as β_k, this order requires that the intrinsic dimension of each component is about $(n/q)^{1/(2r+1)}$, and the total dimension for q components is $q^{2r/(2r+1)}n^{1/(2r+1)}$. Theorem 3 shows that the penalty term with order $q^{2r/(2r+1)}n^{1/(2r+1)}\log(n)$ yields a strong consistent order estimator. This term depends on the decay rate of the eigenvalues of the operator $K^{1/2}CK^{1/2}$. We can compare this with the popular BIC penalty

$$\mathrm{BIC}_n(q)=\frac{dq+q-1}{2}\log(n),$$

where d is the dimension of the parameter space. However, owing to the existence of the infinite-dimensional parameters, choosing the BIC penalty in our setting is not straightforward. Theorems 1–3 together show that the minimax rate of convergence for $\mathcal{R}_n$ is $(q^*/n)^{2r/(2r+1)}$, which is determined by the rate of decay of the eigenvalues of the operator $K^{1/2}CK^{1/2}$ and the model order q^*.

2.2. The estimation algorithm

The estimation procedure for model (1.1) consists of two parts: (i) the estimation of q^*, and (ii) the estimation of the coefficient functions $\Theta^*_{q^*}$. We propose the following two-step estimation procedure. First, for different values of q, we obtain the estimation of Θ_q for each given q. Second, based on the estimation of Θ_q, we compute the term inside the maximum in (2.5), and then determine the estimate of q^* accordingly. In this section, we fix f_0 to be a Gaussian density. With an abuse of notation, we use $f_0(Y_i|X_i,\beta_k,\sigma^2)$ to denote $f_0(Y_i-\langle X_i,\beta_k\rangle)$.

EM algorithm

For a given q, the estimate of Θ_q can be obtained using an EM algorithm. Define u_{ik} as an indicator of whether X_i is from component k, that is, $u_{ik}=1$ if X_i comes from the kth component, and $u_{ik}=0$ otherwise. If the missing data u_{ik} can be observed, then the penalized log-likelihood for the complete data is given by

$$\log L_c(\Theta_q,\sigma^2)=\sum_{i=1}^{n}\sum_{k=1}^{q}u_{ik}\left\{\log\pi_k+\log f_0(Y_i|X_i,\beta_k,\sigma^2)\right\}-\sum_{k=1}^{q}\frac{n\lambda||\beta_k||_K^2}{2\sigma^2},$$

where λ controls the level of the penalty.

The corresponding EM algorithm is given in the following E-step and M-step:

E-step:

Denote the estimates of the parameters in the mth iteration as $\Theta_q^{(m)} = \{\beta_1^{(m)}, \ldots, \beta_q^{(m)}\}$, $\Pi_q^{(m)} = \{\pi_1^{(m)}, \ldots, \pi_q^{(m)}\}$, and $\sigma^{2(m)}$. We estimate $\tau_{ik} = P(u_{ik} = 1)$ as follows:

$$\tau_{ik}^{(m+1)} = \frac{\pi_k^{(m)} f_0\left(Y_i | X_i, \beta_k^{(m)}, \sigma^{2(m)}\right)}{\sum_{k=1}^{q} \pi_k^{(m)} f_0\left(Y_i | X_i, \beta_k^{(m)}, \sigma^{2(m)}\right)}. \tag{2.7}$$

The mixing probabilities $\Pi^{(m+1)}$ are estimated accordingly by

$$\pi_k^{(m+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{(m)}. \tag{2.8}$$

M-step:

We maximize the following Q function with respect to Θ_q and σ^2:

$$Q\left(\Theta_q, \sigma^2\right) = \sum_{i=1}^{n} \sum_{k=1}^{q} \tau_{ik}^{(m)} \left(\log \pi_k + \log f_0\left(Y_i; X_i, \beta_k, \sigma^2\right)\right) - \sum_{k=1}^{q} \frac{n\lambda ||\beta_k||_K^2}{2\sigma^2}. \tag{2.9}$$

By the representer theorem (Wahba (1990)), for each β_k, there exists a vector $C_k = (c_{1k}, \ldots, c_{nk})^T \in R^n$ such that the quantity that maximizes the Q function (2.9) can be expressed as

$$\beta_k(t)^{(m+1)} = \sum_{i=1}^{n} c_{ik} \int_{\mathcal{I}} K(t, s) X_i(s) ds \tag{2.10}$$

and $||\beta_k||_K^2 = C_k^T W C_k$, where $W = (W_{ij})$ is an n-by-n Gram matrix with

$$W_{ij} = \iint_{\mathcal{I} \times \mathcal{I}} X_i(s) K(s, t) X_j(t) ds dt \quad \text{for} \quad i, j \in \{1, \ldots, n\}. \tag{2.11}$$

Therefore, the maximization of the Q function (2.9) can be solved in a componentwise manner by solving the optimization with respect to C_k.

We consider the case when f is a normal distribution. Then the conditional distribution of Y_i is given by

$$Y_i | X_i \sim N\left(\int_{\mathcal{I}} X_i(t) \beta_k(t) dt, \sigma^2\right) = N\left(C_k^T W_{\cdot,i}, \sigma^2\right),$$

where $W_{\cdot,i}$ is the ith column of the Gram matrix (2.11).

The Q function (2.9) then reduces to (up to a constant)

$$Q\left(\Theta_q, \sigma^2\right) = \sum_{i=1}^{n}\sum_{k=1}^{q} \tau_{ik}^{(m)} \left(\log \pi_k - \frac{1}{2}\log \sigma^2 - \frac{\left(Y_i - C_k^T W_{\cdot,i}\right)^2}{2\sigma^2} \right) - \sum_{k=1}^{q} \frac{n\lambda C_k^T W C_k}{2\sigma^2}. \quad (2.12)$$

The estimate of C_k in the above equation can be expressed as a Tikhonov regularized weighted least squares problem. The solution is given by

$$C_k^{(m+1)} = \left(W^T T_k^{(m)} W + n\lambda W\right)^{-1} W^T T_k^{(m)} Y, \quad (2.13)$$

where $T_k^{(m)} = \text{diag}\{\tau_{1k}^{(m)}, \ldots, \tau_{nk}^{(m)}\}$ and $Y = (Y_1, \ldots, Y_n)^T$.

We then construct the coefficient functions β_k by plugging $c_{ik}^{(m+1)}$ into (2.10); that is,

$$\beta_k(t)^{(m+1)} = \sum_{i=1}^{n} c_{ik}^{(m+1)} \int_{\mathcal{I}} K(t,s) X_i(s) ds. \quad (2.14)$$

The estimation of σ^2 is achieved by solving the penalized MLE in (2.12), leading to

$$\sigma^{2(m+1)} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{q} \tau_{ik}^{(m)} \left(Y_i - C_k^{T(m+1)} W_{\cdot,i}\right)^2}{\sum_{i=1}^{n}\sum_{k=1}^{q} \tau_{ik}^{(m)}} + \frac{n\lambda \sum_{k=1}^{q} C_k^T W C_k}{\sum_{i=1}^{n}\sum_{k=1}^{q} \tau_{ik}^{(m)}}. \quad (2.15)$$

We iteratively update (2.7) and (2.8) in the E-step and (2.13), (2.14), and (2.15) in the M-step until the convergence is reached.

We use the Gaussian kernel throughout this paper. The bandwidth, denoted by ϱ, of the reproducing kernel $K(\cdot,\cdot)$ and the ridge penalty λ are treated as hyperparameters. They are tuned using cross-validation. Other kernels can be easily incorporated into our algorithm. Learning the optimal kernels is a nontrivial matter, and deserves further investigation. A good method of doing so is to combine multiple candidate kernels.

Order estimation

The optimal order of the mixture model can be estimated using the optimization in (2.5). This is a mixed integer programming (MIP) problem, the solution of which may not be obtained in practice. We denote the negative of Equation (2.5) inside the maximum as the order objective function, that is,

$$\text{Order}(q) = \Big(-\ell_n(f^{(q)}) + \lambda \|\beta^{(q)}\|_K^2 \Big) + \text{pen}_n(q), \quad (2.16)$$

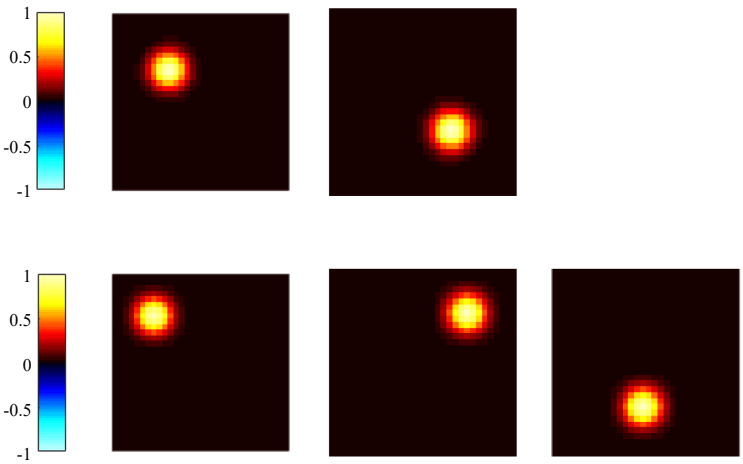

Figure 1. The first row of images includes the true coefficients β_1 and β_2 in the first scenario, and the bottom row includes the true coefficients β_1, β_2, and β_3 in the second scenario.

where $f^{(q)}$ and $\beta^{(q)}$ are the estimates calculated from the EM algorithm with the model order being q. The first part of Order(q) is essentially the penalized loss, and the second part is the penalty (2.6) on q. Thus, we conduct a sequential search for the optimal order q^*. Specifically, we start from $q = 1$ and increase its value by one each time, and then compute Order(q) using (2.16). As the model order increases, the penalized loss part in Order(q) decrease, and the penalty on the model order increases. Once Order(q) no longer decreases, we select the q value that minimizes Order(q) as the optimal estimate.

3. Numerical Results

We conduct simulations to evaluate the accuracy, and efficiency of our methods and then apply them to an analysis of a real-data set obtained from the ADNI study.

3.1. Synthetic experiments

We perform simulations in two different scenarios, including one with two mixing components ($q^* = 2$) and another with three ($q^* = 3$). The true coefficient images are generated from a Gaussian function with different centers. The covariate images $X_i(t)$ are simulated from a Gaussian random field. Figure 1 displays the true coefficients for these two scenarios.

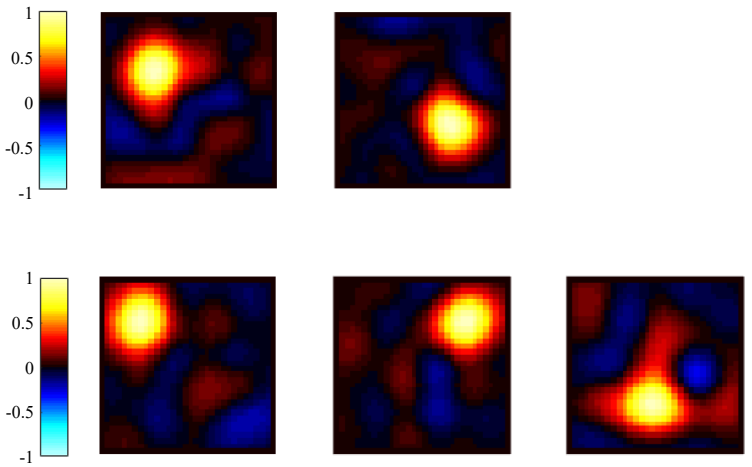

Figure 2. The top row of images includes estimates of β_1 and β_2 in scenario 1 from one replication, and the bottom row contains estimates of β_1, β_2, and β_3 in scenario 2.

The responses Y_i follow a mixture of functional linear models given by

$$Y_i = \sum_{k=1}^{q} \delta_{ik} \left(\beta_{k0} + \int_{t \in \mathcal{T}} X_i(t) \beta_k(t) dt \right) + \epsilon_i, \tag{3.1}$$

where $\delta_{ik} = 1$ if subject i comes from the kth component, and zero otherwise. The noise term ϵ_i follows a standard normal distribution and is independent of $X_i(t)$. In each scenario, we set the sample size of each component to be 80, that is, $\sum_{i=1}^{n} \delta_{ik} = 80$, and the total number of samples is $n = 160$ and 240 in scenarios 1 and 2, respectively.

The initial tuning range for the bandwidth ϱ of the kernel K and the ridge penalty level λ is from a two-dimensional grid, that is, $\sigma \otimes \lambda = \{2^{-5}, 2^{-4}, \ldots, 2^5\}^{\otimes 2}$. We start with the functional linear model with one component and increase the value q by one each time. For each q, the stopping criterion of the EM algorithm is either determined by $\sum_{k=1}^{q} \|C_k^{(m+1)} - C_k^{(m)}\|^2 / \{\sum_{k=1}^{q} \|C_k^{(m)}\|^2\} \leq 10^{-5}$ or is a total number of iterations over 500. To evaluate the estimation performance, we first check the estimation of q^*. If q^* is correctly estimated, we further compute the relative mean squared error (RMSE) for each β_k, that is, RMSE $= \|\widehat{\beta}_k - \beta_k\|^2 / \|\beta_k\|^2$. For each sample, we denote $\widehat{u}_{ik} = 1$ if $k = \arg\max_j \{\widehat{\tau}_{ij}\}$, and compute the assignment accuracy as ACC $= n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{q^*} \delta(\widehat{u}_{ik}, u_{ik})$, where $\delta(\cdot, \cdot)$ is the indicator function. The empirical prediction risk is quantified as the root mean square prediction error, that is, RMSPE $= \sqrt{(1/n) \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}$. We also conduct a regression analysis using just the functional linear model (FLM)

Table 1. The empirical mean and standard error (in parenthesis) of the relative mean squared error (RMSE), assignment accuracy (ACC), and root mean square prediction error (RMSPE) in scenarios 1 and 2, labeled as S1 and S2, respectively. FFMRM denotes the functional finite mixture regression model; FLM denotes the functional linear model. For each case, 100 simulated data sets were used.

	RMSE			**ACC**	**RMSPE**	
	β_1	β_2	β_3		FFMRM	FLM
S1	1.80(0.38)	1.84(0.39)	N/A	0.86(0.06)	1.01(0.47)	3.65(0.35)
S2	1.99(0.39)	2.04(0.37)	2.10(0.28)	0.65(0.23)	1.02(0.34)	3.53(0.32)

Table 2. The demographic information of all participants in the ADNI study. The means are reported, with the standard deviations included in parentheses.

Diagnosis	Gender(F/M)	Age (years)	MMSE
CN	164/167	74.46 (5.58)	28.99 (1.23)
SMC	61/42	72.01 (5.46)	29.01 (1.96)
EMCI	127/161	71.26 (7.51)	28.35 (1.55)
LMCI	152/245	73.66 (7.41)	26.97 (2.70)
AD	99/142	74.96 (7.88)	23.25 (2.10)

(without identifying the mixing components) as a baseline for the RMSPE comparison. Note that the basic FLM only has one component, so the RMSE and ACC comparisons are not applicable.

We present the estimation results in each scenario in Table 1. In both scenarios, we can correctly estimate the order of the mixing components q^* in all 100 repetitions. In scenario 1, the assignment accuracy is 86.5%, whereas in scenario 2, it is 65.1%. This decrease of accuracy is mainly due to the increase of the number of components.

We achieve desirable RMSEs in both scenarios, with slightly better performance in scenario 1. To graphically demonstrate the estimated coefficient, we randomly pick up one iteration for each scenario and plot the estimated coefficients in Figure 2. The estimated coefficients correctly capture the nonzero regions and recover the patterns in the true coefficients.

3.2. Mini-mental state examination score prediction in the ADNI study

To further demonstrate the usefulness of model (1.1), we apply our methods to a real-data set obtained from ADNI study. The ADNI study is a large scale multi-site study and has collected magnetic resonance imaging (MRI) images, positron emission tomography (PET) images, cerebrospinal fluid (CSF), and blood biomarkers, among many others. More information about this study

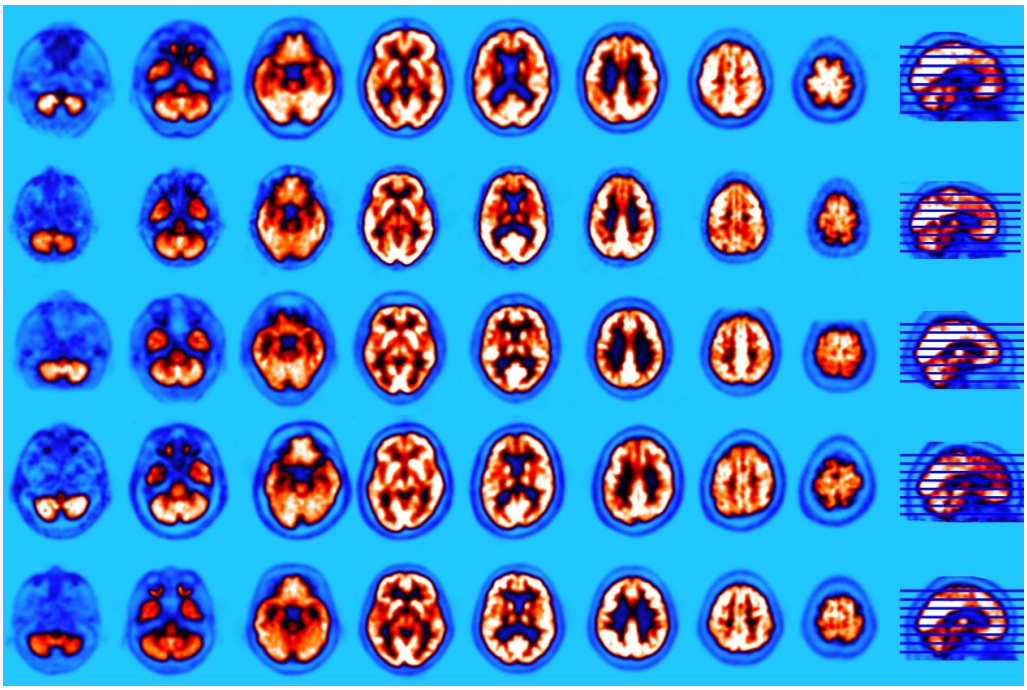

Figure 3. The five rows of images are the PET images of different types of participants. We randomly pick one image from each group for illustration. The types of participants are AD, CN, EMCI, LMCI, and SMC, from top to bottom, respectively. Eight transverse slices located at $Z = 8$ to 78 with equal increments of 10 are displayed from left to right.

can be found at the ADNI website (`http://adni.loni.usc.edu/`).

We study the PET data, and use them as imaging predictors to predict the mini-mental state examination (MMSE) scores. The PET images measure the metabolic processes of the patients, such as flows of blood to different parts of the brain, by detecting the radioactivity of the injected tracer. We analyze the PET images collected at the baseline of the ADNI study. Among the 1,360 participants, there are five diagnostic groups: normal control (CN), significant memory concern (SMC), early mild cognitive impairment (EMCI), late mild cognitive impairment (LMCI) and those diagnosed with Alzheimer's disease (AD). The demographic information of the participants is summarized in Table 2, and a sample image from each type of participant is provided in Figure 3.

We remove the age and gender factors by fitting a linear model in each voxel, and use the residual maps as the covariate images of model (1.1). The same configuration of the EM algorithm in the simulation studies is used to train the model, and the adaptive tuning procedure is used to tune the parameters. Finally, three mixing components are determined by the EM algorithm, with their

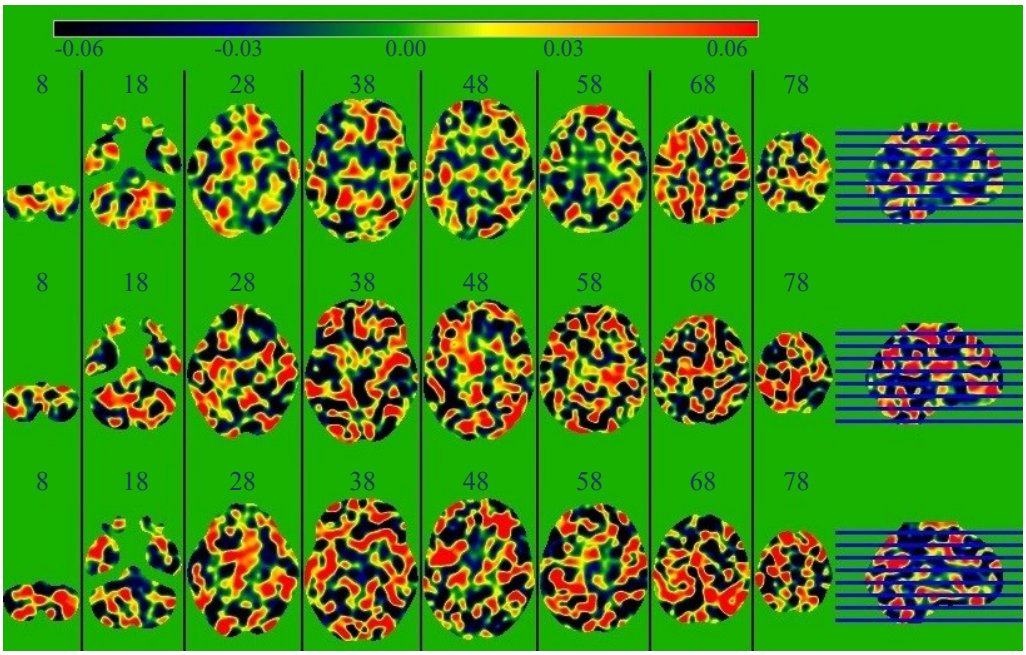

Figure 4. The three rows of images are the estimated coefficients $\widehat{\beta}_1$, $\widehat{\beta}_2$, and $\widehat{\beta}_3$, from top to bottom, respectively. Eight transverse slices located at $Z = 8$ to 78 with equal increments of 10, are displayed from left to right.

corresponding coefficient images displayed in Figure 4. The overall proportional probabilities π_k are estimated as 19.8%, 41.0%, and 39.2%, respectively. We conduct a hierarchical clustering on the patients according to their estimated proportional probabilities to each component τ_{ik}, and illustrate the corresponding similarity matrix in Figure 5. These clusters indicate potential heterogeneous disease patterns of AD among the population, which has been proved in previous studies (Latta, Brothers and Wilcock (2015); Dong et al. (2016)). We overlay the coefficient images on the MNI-152 ROI template (Fonov et al. (2011)), and identify several regions of interest: $\widehat{\beta}_1$ mainly represents impairment in the sup frontal and parietal lobe; $\widehat{\beta}_2$ highlights the corpus callosum and occipital lobe; and $\widehat{\beta}_3$ mostly covers the middle frontal gyrus. Many prior studies have shown that reduced glucose metabolic activities and structural impairment in the frontal lobe, parietal lobe, and occipital lobe of the cerebrum are associated with a degradation in cognition and a progression to AD (Lin et al. (2017); Heneka et al. (2015)).

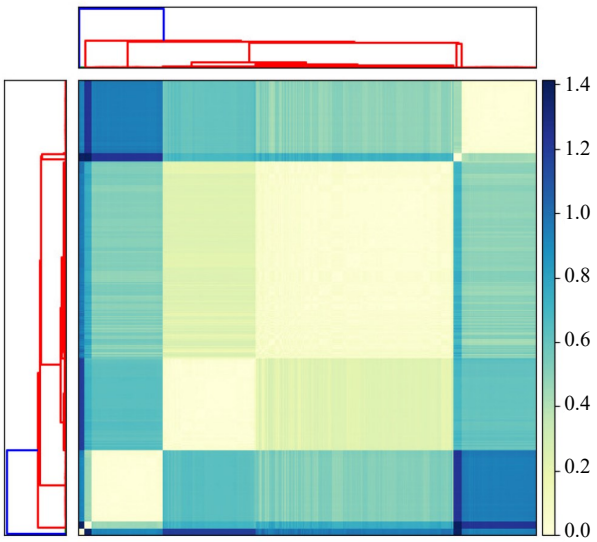

Figure 5. The similarity matrix of the estimation of τ_k. The value indicates the pairwise Euclidean distance of the proportional probabilities.

4. Conclusion

We have developed a functional finite mixture regression model with functional predictors in the RKHS. This is a challenging problem because of the unknown model order and the existence of infinite-dimensional unknown parameters. Our theoretical developments provide a strong consistent estimator of the model order using a penalized likelihood estimation and establish the minimax rate of convergence for the estimation risk. We have shown that the optimal rate is jointly determined by the alignment of the reproducing kernel, covariance kernel, and model order. An efficient EM algorithm is also proposed, and empirical experimental results demonstrate the merits of our method.

Appedix: Technical Lemmas and Proofs of Theorems

A1. Technical Lemmas

Lemma 1. *Assume* B1 *and* B2 *hold. Define* $H_q(\epsilon) = \{\sqrt{f/f^*} : f \in \mathcal{M}_q, \mathcal{H}(h, h^*) \leq \epsilon\}$. *Then*

$$N_{[\,]}\Big(H_q(\epsilon), \delta\Big) \leq \left(\frac{C_1\ \epsilon}{\delta}\right)^{10\left(C_2\delta^{-1/r}+1\right)q+1} \tag{A.1}$$

for all $q \geq q^*$ *and* $\delta/\epsilon \leq 1$.

Proof. We prove the lemma by extending the arguments in Gassiat and van Handel (2014). Note that $f/f^* = h/h^*$. We denote by $\|\cdot\|_p$ the $L^p(h^* d\mu)$-norm, that is, $\|g\|_p^0 = \int |g|^p h^* d\mu$. Denote $\eta_i^* = \langle x, \beta_i^* \rangle$ and $\eta_j = \langle x, \beta_j \rangle$, $i = 1, \ldots, q^*$, $j = 1, \ldots, q$. It follows from B2 that we can find a partition of $\mathbb{R}$, $A_0, A_1, \ldots, A_{q^*}$, such that each bounded set $A_i, i = 1, \ldots, q^*$, contains precisely one component $\eta_i^* = \langle X, \beta_i^* \rangle$ and the unbounded set $A_0 = \mathbb{R}^M \setminus (A_1 \cup \ldots \cup A_{q^*})$ contains no component. Let $f \in \mathcal{M}_q$, so that we can write $f = \sum_{i=1}^q \pi_i f_0(y - \eta_i)$. Then,

$$\frac{f - f^*}{f^*} = \sum_{j:\eta_j \in A_0} \pi_j \frac{f_0(y - \eta_j)}{f^*} + \sum_{i=1}^{q^*} \left\{ \left(\sum_{j:\eta_j \in A_i} \pi_j - \pi_i^* \right) \frac{f_0(y - \eta_i^*)}{f^*} + \sum_{j:\eta_j \in A_i} \frac{f_0(y - \eta_j) - f_0(y - \eta_i^*)}{f^*} \right\}.$$

Taylor expansion gives

$$f_0(y - \eta_j) - f_0(y - \eta_i^*) = \dot{f}_0(y - \eta_i^*)(\eta_j - \eta_i^*) + \frac{1}{2} \ddot{f}_0(y - \tilde{\eta})(\eta_j - \eta_i^*)^2.$$

Using Assumption B2, we find that

$$\left| \frac{f - f^*}{f^*} \right| \leq \left[\sum_{j:\eta_j \in A_0} \pi_j + \sum_{i=1}^{q^*} \left\{ \left| \sum_{j:\eta_j \in A_i} \pi_j - \pi_i^* \right| + \left| \sum_{j:\eta_j \in A_i} \pi_j (\eta_j - \eta_i^*) \right| + \frac{1}{2} \sum_{j:\eta_j \in A_i} \pi_j (\eta_j - \eta_i^*)^2 \right\} \right] (H_0 + H_1 + H_2).$$

On the other hand, it follows from Theorem 3.10 of Gassiat and van Handel (2014) that there exists a constant c^* such that

$$\left\| \frac{f - f^*}{f^*} \right\|_1 \geq c^* \left[\sum_{j:\eta_j \in A_0} \pi_j + \sum_{i=1}^{q^*} \left\{ \left| \sum_{j:\eta_j \in A_i} \pi_j - \pi_i^* \right| + \left| \sum_{j:\eta_j \in A_i} \pi_j (\eta_j - \eta_i^*) \right| + \frac{1}{2} \sum_{j:\eta_j \in A_i} \pi_j (\eta_j - \eta_i^*)^2 \right\} \right].$$

Hence, for all $f \in \mathcal{M}$,

$$\frac{|f/f^* - 1|}{\|f/f^* - 1\|_1} \leq S := \frac{1}{c^*} (H_0 + H_1 + H_2).$$

In addition, using $|\sqrt{x}-1| \le |x-1|$, we find

$$\frac{|\sqrt{h/h^*}-1|}{\mathcal{H}(h,h^*)} \le \frac{|h/h^*-1|}{(1/2)\|h/h^*-1\|_1} = \frac{|f/f^*-1|}{(1/2)\|f/f^*-1\|_1} \le 2S.$$

Similar to Lemma 3.15 of [1], for and $f \in \mathcal{M}$, we have

$$\frac{|\sqrt{h/h^*}-1|}{H(h,h^*)} - \frac{h/h^*-1}{\sqrt{\chi^2(h\|h^*)}} \le (4\|S\|_4^2 S + 2S^2)H(h,h^*),$$

where the chi-square divergence is defined as $\chi^2(h\|h^*) = \int (h/h^*-1)^2 h^* d\mu$. This allows us to make further approximation based on Lemma 3.16 of [1]. Let $\alpha > 0$, and for every $f \in \mathcal{M}_q$ such that $H(h,h^*) \le \alpha$. Define

$$\tilde{\ell} = \sum_{i=1}^{q^*} \left\{ a_i \frac{f_0(y-\eta_i^*)}{f^*} + b_i \frac{\dot{f}_0(y-\eta_i^*)}{f^*} + e_i^2 \frac{\ddot{f}_0(y-\eta_i^*)}{f^*} \right\} + \sum_{j=1}^{q} \gamma_j \frac{f_0(y-\eta_j)}{f^*},$$

where

$$\begin{aligned} &\sum_{i=1}^{q^*} |a_i| \le \frac{1}{c^*} + \frac{1}{\sqrt{c^*\alpha}}, \quad \sum_{i=1}^{q^*} |b_i| \le \frac{1}{c^*} + 2\frac{c}{\sqrt{c^*\alpha}}, \\ &\sum_{i=1}^{q^*} e_i^2 \le \frac{1}{c^*}, \quad \sum_{j=1}^{q} \gamma_j \le \frac{1}{\sqrt{c^*\alpha} \wedge c^*}. \end{aligned} \tag{A.2}$$

We have

$$\Big| \frac{h/h^*-1}{\sqrt{\chi^2(h\|h^*)}} - \tilde{\ell} \Big| \le \frac{\sqrt{2}}{2(c^*)^{5/4}} (\|H_3\|_2 S + H_3)\alpha^{1/4}.$$

Define

$$d_f = \frac{\sqrt{f/f^*}-1}{\|\sqrt{f/f^*}-1\|_2},$$

and

$$\mathcal{D}_q = \{d_f : f \in \mathcal{M}_q, f \ne f^*\}, \quad \mathcal{D}_{q,\alpha} = \{d_f : f \in \mathcal{M}_q, f \ne f^*, \mathcal{H}(h,h^*) \le \alpha\}.$$

Then,

$$N_{[\,]}(\mathcal{D}_q, \delta) \le N_{[\,]}(\mathcal{D}_{q,\alpha}, \delta) + N_{[\,]}(\mathcal{D}_q \setminus \mathcal{D}_{q,\alpha}, \delta).$$

We estimate both bracketing numbers separately. Define a family of functions

$$\tilde{\mathcal{L}}_{q,\alpha} = \left\{ \sum_{i=1}^{q^*} \left\{ a_i \frac{f_0(y-\eta_i^*)}{f^*} + b_i \frac{\dot{f}_0(y-\eta_i^*)}{f^*} + e_i^2 \frac{\ddot{f}_0(y-\eta_i^*)}{f^*} \right\} + \sum_{j=1}^{q} \gamma_j \frac{f_0(y-\eta_j)}{f^*} \right\},$$

where $(a, b, e, \gamma) \in \mathbb{R}^{q^*} \times R^{q^*} \times R^{q^*} \times R^q$ satisfies the constraint (A.2). For $\tilde{l}, \tilde{l}' \in \tilde{\mathcal{L}}_{q,\alpha}$,

$$\begin{aligned} |\tilde{l} - \tilde{l}'| \le H_0 \sum_{i=1}^{q^*} |a_i - a_i'| + H_1 \sum_{i=1}^{q^*} |b_i - b_i'| + H_0 \sum_{i=1}^{q} |\gamma_i - \gamma_i'| \\ + \frac{1}{\sqrt{c^*\alpha} \wedge c^*} H_1 \sum_{i=1}^{q} |\langle x, \beta_i - \beta_i' \rangle| + \frac{2H_2}{\sqrt{c^*}} \sum_{i=1}^{q^*} |e_i - e_i'|. \end{aligned}$$

Note that

$$\begin{aligned} \mathbb{E}_X \left(\sum_{i=1}^{q} |\langle x, \beta_i - \beta_i' \rangle| \right)^2 &\le q \sum_{i=1}^{q} \mathbb{E}_X |\langle x, \beta_i - \beta_i' \rangle|^2 = q \sum_{i=1}^{q} \sum_{k=1}^{\infty} \rho_k (g_{ik} - g_{ik}')^2 \\ &\le q \sum_{i=1}^{q} \sum_{k=1}^{M} \rho_k (g_{ik} - g_{ik}')^2 + M^{-2r} o_M(1). \end{aligned}$$

We may choose $M = c_1 \delta^{-1/r}$, so that

$$|\tilde{l} - \tilde{l}'| \le V \|(a, b, \gamma, g, e) - (a', b', \gamma', g', e')\|_{\text{norm}} + o(\delta),$$

where $V = 3(H_0 + H_1 + H_2)$ and $\|\cdot\|_{\text{norm}}$ is the norm on $\mathbb{R}^{3q^* + c_2 q(\delta^{-1/r}+1)}$ defined by

$$\begin{aligned} \|(a, b, \gamma, g, e)\|_{\text{norm}} = \sum_{i=1}^{q^*} |a_i| + \sum_{i=1}^{q^*} |b_i| + \sum_{i=1}^{q} |\gamma_i| \\ + \frac{q}{\sqrt{c^*\alpha} \wedge c^*} \sum_{i=1}^{q} \sum_{k=1}^{M} \rho_k g_{ik}^2 + \frac{2}{\sqrt{c^*}} \sum_{i=1}^{q^*} |e_i|. \end{aligned}$$

Using the standard fact of the covering number for the Euclidean ball we obtain

$$N_{[\,]}\left(\tilde{L}_{q,\alpha}, \delta\right) = \left(\frac{c_3 + \delta}{\delta}\right)^{3q^* + c_2 q(\delta^{-1/r}+1)}.$$

Since $q \ge q^*$ and $\delta \le 1$, we therefore obtain

$$N_{[\,]}\left(\mathcal{D}_{q,\alpha}, \delta\right) \le \left(\frac{c_4}{\delta}\right)^{c_5 q(\delta^{-1/r}+1)}.$$

Lemma 2. *Assume* B1 *and* B2 *hold.*

(a).

$$\mathbb{P}\left[\sup_{f\in\mathcal{M}_q}\sum_{i=1}^{n}\log\frac{f(Y_i|X_i)}{f^*(Y_i|X_i)}\geq\alpha\right]\leq C_3e^{-\alpha/C_3}$$

for all $\alpha\geq C_4q^{2r/(2r+1)}n^{1/(2r+1)}$.

(b).

$$\overline{\lim_{n\to\infty}}\sup_{f\in\mathcal{M}_q}\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(Y_i|X_i)}{f^*(Y_i|X_i)}<0,\ \ a.s.$$

Proof. Let $h(y,x)=f(y|x)f_X(x)$ and $h^*(y,x)=f^*(y|x)f_X(x)$ for $f\in\mathcal{M}_q$ and $f^*\in\mathcal{M}_{q^*}$. Denote by $\bar{h}=(h+h^*)/2$. We first have the inequalities $\mathcal{K}(h,h^*)\geq\mathcal{H}^2(h,h^*)$, and

$$\sum_{i=1}^{n}\log\frac{h}{h^*}\leq 2\sqrt{n}\nu_n\bigg(\log\frac{\bar{h}}{h^*}\bigg)-2n\mathcal{K}(\bar{h},h^*). \tag{A.3}$$

Therefore,

$$\begin{aligned}\mathbb{P}\bigg[\sup_h\sum_{i=1}^{n}\log\frac{h}{h^*}\geq\alpha\bigg]&\leq\mathbb{P}\bigg[\sup_h\sqrt{n}\nu_n\bigg(\log\frac{\bar{h}}{h^*}\bigg)-n\mathcal{H}^2(\bar{h},h^*)\geq\frac{\alpha}{2}\bigg]\\&\leq\sum_{s=0}^{S}\mathbb{P}\bigg[\sup_{h\in\mathcal{G}_s}\nu_n\bigg(\log\frac{\bar{h}}{h^*}\bigg)\geq\frac{\alpha2^{s-1}}{\sqrt{n}}\bigg]\\&=\sum_{s=0}^{S}\mathbb{P}\bigg[\sup_{h\in\mathcal{G}_s}\nu_n\bigg(\log\sqrt{\frac{\bar{h}}{h^*}}\ \bigg)\geq\frac{\alpha2^{s-2}}{\sqrt{n}}\bigg],\end{aligned}$$

where $\mathcal{G}_0=\{\bar{h}:n\mathcal{H}^2(\bar{h},h^*)\leq\alpha\}$, $\mathcal{G}_s=\{\bar{h}:\alpha2^{s-1}<n\mathcal{H}^2(\bar{h},h^*)\leq\alpha2^s\}$, $1\leq s\leq S$, $S=\min\{s:\alpha2^s>2n\}$. We need to find the bracketing number for $\bar{H}_q(\epsilon)=\{\sqrt{\bar{h}/h^*}:\mathcal{H}(\bar{h},h^*)\leq\epsilon\}$. This can be easily deduced from Lemma 1 such that

$$N_{[\ ]}\Big(\bar{H}_q(\epsilon),\delta\Big)\leq\left(\frac{2\sqrt{2}C_1\ \epsilon}{\delta}\right)^{10\left(C_2\delta^{-1/r}+1\right)q+1}. \tag{A.4}$$

Note that $\int_0^\epsilon\sqrt{\log N_{[\ ]}(\bar{H}_q(\epsilon),\delta)}d\delta\leq c\sqrt{q}\epsilon^{1-1/(2r)}$. It requires that $c\sqrt{q}\epsilon^{1-1/(2r)}\leq\sqrt{n}\epsilon^2$ such as $\epsilon\geq c_1(q/n)^{r/(2r+1)}$. Next, apply Theorem 7.4 of van de Geer (2000), together with $\sqrt{\alpha}2^{s/2+2}/\sqrt{n}\geq c_1(q/n)^{r/(2r+1)}$, that is, $\alpha\geq c_2q^{2r/(2r+1)}n^{1/(2r+1)}$.

We obtain

$$\sum_{s=0}^{S}\mathbb{P}\left[\sup_{h\in\mathcal{G}_s}\nu_n\left(\log\sqrt{\frac{\bar{h}}{h^*}}\right)\geq\frac{\alpha 2^{s-2}}{\sqrt{n}}\right]\leq\sum_{s=0}^{S}c_3e^{-\alpha 2^s/(c_3 2^8)}\leq C_3e^{-\alpha/C_3}.$$

This finishes the proof of Part (a).

To prove Part (b), it follows from (A.3) that it is enough to show that

$$\lim_{n\to\infty}\sup_{f\in\mathcal{M}_q}n^{-1/2}\nu_n\left(\log\sqrt{\frac{\bar{h}}{h^*}}\right)=0,\quad\text{a.s.}$$

As in the proof of Part (a), we have

$$\mathbb{P}\left[\sup_{f\in\mathcal{M}_q}n^{-1/2}\nu_n\left(\log\sqrt{\frac{\bar{h}}{h^*}}\right)\geq\alpha\right]\leq c_4e^{-n\alpha^2/c_4}$$

for every $\alpha>0$ such that $c_2q^{2r/(2r+1)}n^{1/(2r+1)}\leq\alpha\sqrt{n}\leq 32\sqrt{n}$. Hence,

$$\sum_{n=1}^{\infty}\mathbb{P}\left[\sup_{f\in\mathcal{M}_q}n^{-1/2}\nu_n\left(\log\sqrt{\frac{\bar{h}}{h^*}}\right)\geq\alpha\right]\leq\infty$$

for $0<\alpha\leq 32$. Part b follows from Borel-Cantelli.

Lemma 3. *Assume* B1 *and* B2 *hold. Define*

$$\Delta_n(q,q^*)=\sup_{f\in\mathcal{M}_q}\left(\ell_n(f)-\lambda_{q,n}\|\beta\|_K^2\right)-\sup_{f\in\mathcal{M}_{q^*}}\left(\ell_n(f)-\lambda_{q^*,n}\|\beta\|_K^2\right),$$

where $\lambda_{q,n}=C_5q^{2r/(2r+1)}n^{1/(2r+1)}$ *and* $\lambda_{q^*,n}=\tilde{C}_5(q^*)^{2r/(2r+1)}n^{1/(2r+1)}$. *Then,*

$$\overline{\lim_{n\to\infty}}\sup_{q>q^*}\frac{\Delta_n(q,q^*)}{q^{2r/(2r+1)}n^{1/(2r+1)}}\leq C_6,\quad a.s.\tag{A.5}$$

Proof. Using the fact that

$$\sup_{f\in\mathcal{M}_{q^*}}\left(\ell_n(f)-\lambda_{q^*,n}\|\beta\|_K^2\right)\geq\ell_n(f^*)-\lambda_{q^*,n}\|\beta^*\|_K^2$$

yield

$$\Delta_n(q,q^*)\leq\sup_{f\in\mathcal{M}_q}\sum_{i=1}^{n}\log\frac{f(Y_i|X_i)}{f^*(Y_i|X_i)}+\lambda_{q^*,n}\|\beta^*\|_K^2.$$

Furthermore, from Part (a) of Lemma 2, for $\alpha\geq C_4$,

$$\mathbb{P}\left[\sup_{q\geq q^*}\frac{1}{q^{2r/(2r+1)}n^{1/(2r+1)}}\sup_{f\in\mathcal{M}_q}\sum_{i=1}^{n}\log\frac{f(Y_i|X_i)}{f^*(Y_i|X_i)}\geq\alpha\right]$$
$$\leq\sum_{q=q^*}^{\infty}C_3e^{-\alpha q^{2r/(2r+1)}n^{1/(2r+1)}/C_3}\leq\frac{c_1}{n^2}.$$

The lemma follows easily using the Borel-Cantelli.

A2. Proof of Theorem 1

The constants c_i used in the proof are all generic positive constants. First realize that any lower bound for a specific case yields immediately a lower bound for the general case. In the following, consider a special case where q^* and $(\pi_1^*,\ldots,\pi_{q^*}^*)$ are known. f_0 is a fixed known density function.

Direct calculation of the Kullback-Leibler divergence between $h^*=f^*(y|x)$ $f_X(x)$ and $h=f(y|x)f_X(x)$ where $f^*(y|x), f(y|x)\in\mathcal{M}_{q^*}$ yields

$$\mathcal{K}(h,h^*)=\mathbb{E}_{h^*}\left\{\log\frac{h^*}{h}\right\}=\mathbb{E}_{h^*}\left\{\log\frac{\sum_{k=1}^{q^*}\pi_k^*h_k^*}{\sum_{k=1}^{q^*}\pi_k^*h_k}\right\}=\mathbb{E}_{h^*}\left\{\log\sum_{k=1}^{q^*}\tau_k\frac{h_k^*}{h_k}\right\}$$
$$\geq\mathbb{E}_{h^*}\left\{\sum_{k=1}^{q^*}\tau_k\log\frac{h_k^*}{h_k}\right\}=\sum_{k=1}^{q^*}\pi_k^*\mathbb{E}_{h_k^*}\left\{\frac{h_k/h}{h_k^*/h^*}\log\frac{h_k^*}{h_k}\right\},\tag{A.1}$$

where $\tau_k=\pi_k^*h_k/h$ for $k=1,\ldots,q^*$ and the inequality is due to the Jensen's Inequality. Note that $\pi_k^*h_k/h=\pi_k^*f_0(y-\eta_k)/\sum_{i=1}^{q^*}\pi_i^*f_0(y-\eta_i)$ is the probability of a data point coming from the kth component. We assume this probability bound below away from zero. Therefore, h_k/h is bounded below away from zero and above by a positive constant. Hence,

$$\mathcal{K}(h,h^*)=\sum_{k=1}^{q^*}\pi_k^*\mathbb{E}_{h_k^*}\left\{\frac{h_k/h}{h_k^*/h^*}\log\frac{h_k^*}{h_k}\right\}$$
$$=\sum_{k=1}^{q^*}\pi_k^*\int_{h_k^*>h_k}\frac{h_k/h}{h_k^*/h^*}h_k^*\log\frac{h_k^*}{h_k}d\mu+\sum_{k=1}^{q^*}\pi_k^*\int_{h_k^*\leq h_k}\frac{h_k/h}{h_k^*/h^*}h_k^*\log\frac{h_k^*}{h_k}d\mu,$$
$$\geq c_3\sum_{k=1}^{q^*}\pi_k^*\int_{h_k^*>h_k}h_k^*\log\frac{h_k^*}{h_k}d\mu+c_4\sum_{k=1}^{q^*}\pi_k^*\int_{h_k^*\leq h_k}h_k^*\log\frac{h_k^*}{h_k}d\mu$$
$$\geq c_5\sum_{k=1}^{q^*}\pi_k^*\mathcal{K}(h_k,h_k^*).$$

Using the fact $\log y - \log x - (1/x)(y-x) + (1/(2c_6))(y-x)^2 \le 0$ for $0 \le y \le \sqrt{c_6}$,

$$\begin{aligned}
\mathcal{K}(h_k, h_k^*) &= \mathbb{E}_{h_k^*}\left\{ \log \frac{h_k^*}{h_k} \right\} \\
&\ge \frac{1}{2c_6}\mathbb{E}_{h_k^*}\left\{ f_0(Y - \eta(X, \beta_k)) - f_0(Y - \eta(X, \beta_k^*)) \right\}^2 \\
&= \frac{1}{2c_6}\mathbb{E}_{h_k^*}\left\{ \dot{f}_0{}^2(Y - \tilde{\eta}_k)(\eta(X, \beta_k - \beta_k^*))^2 \right\} \\
&= \frac{1}{2c_6}\mathbb{E}_X\left\{ \mathbb{E}_{f_{0k}(y|x)}(\dot{f}_0{}^2(Y - \tilde{\eta}_k)|X)(\eta(X, \beta_k - \beta_k^*))^2 \right\} \\
&\ge c_7 \mathbb{E}_X\left\{ (\eta(X, \beta_k - \beta_k^*))^2 \right\} \\
&= c_7 \|\beta_k - \beta_k^*\|_C^2,
\end{aligned}$$

where $\tilde{\eta}_k$ is a point between $\eta(X, \beta_k^*)$ and $\eta(X, \beta_k)$. Therefore, $\mathcal{K}(h, h^*)$ is bounded below by $\|\beta - \beta^*\|_C^2$ up to a constant, where $\|\beta - \beta^*\|_C^2 = \sum_{k=1}^{q^*} \|\beta_k - \beta_k^*\|_C^2$. A similar calculation also yields that $\mathcal{K}(h, h^*)$ is bounded above by $\|\beta - \beta^*\|_C^2$ up to a constant.

In the following, we adopt the results from Tsybakov (2009) to establish the lower bound, which is based upon testing multiple hypotheses. In particular, we can find a subset $\{\beta^{(0)}, \ldots, \beta^{(N)}\} \subset \mathcal{F}^{q^*}$ with N increasing with n, such that for some positive constant c and all $0 \le i < j \le N$,

$$\|\beta^{(i)} - \beta^{(j)}\|_C^2 \ge 2c\gamma^{2r/(2r+1)}\left(\frac{n}{q^*}\right)^{-2r/(2r+1)}, \tag{A.2}$$

and

$$\frac{1}{N}\sum_{j=1}^{N} \mathcal{K}(h^{(j)}, h^{(0)}) \le \gamma \log N, \tag{A.3}$$

then we can conclude according to Theorem 2.5 of Tsybakov (2009) that,

$$\inf_{\hat{\beta}} \sup_{\beta^* \in \mathcal{F}^{q^*}} \mathbb{P}\left(\|\beta^{(i)} - \beta^{(j)}\|_C^2 \ge c\gamma^{2r/(2r+1)}\left(\frac{n}{q^*}\right)^{-2r/(2r+1)} \right) \tag{A.4}$$

$$\ge \frac{\sqrt{N}}{1+\sqrt{N}}\left(1 - 2\gamma - \sqrt{\frac{2\gamma}{\log N}}\right), \tag{A.5}$$

which yields

$$\lim_{a\to 0}\lim_{n\to\infty}\inf_{\hat{\beta}}\sup_{\beta^*\in\mathcal{F}^{q^*}}\mathbb{P}\left(\|\hat{\beta}-\beta^*\|_C \geq a\left(\frac{n}{q^*}\right)^{-r/(2r+1)}\right)\geq 1.$$

Hence Theorem 1 will be proved.

Next, we construct the subset $\{\beta^{(0)},\ldots,\beta^{(N)}\}\subset\mathcal{F}^{q^*}$, $k=1,\ldots,q^*$. Let $\beta^{(j)}=(\beta_1^{(j)},\beta_2^{(j)},\ldots,\beta_{q^*}^{(j)})$, $j=1,\ldots,N$. We show that both (A.2) and (A.3) are satisfied. Let $\widetilde{M}=\lfloor M/q^*\rfloor$ for some large number M to be decided later. Consider the function space

$$\mathcal{H}^*=\left\{\beta=\sum_{k=\widetilde{M}+1}^{2\widetilde{M}} b_k M^{-1/2}L_{K^{1/2}\varphi_k}:(b_{M+1},\ldots b_{2M})\in\{0,1\}^M\right\},\qquad(A.6)$$

where $\{\varphi_k:k\geq 1\}$ are the orthonomal eigenfunctions of $K^{1/2}CK^{1/2}$. For any $\beta\in\mathcal{H}^*$, observe that

$$\begin{aligned}||\beta||_K^2&=||\sum_{k=\widetilde{M}+1}^{2\widetilde{M}} b_k M^{-1/2}L_{K^{1/2}\varphi_k}||_K^2\\&=\sum_{k=\widetilde{M}+1}^{2\widetilde{M}} b_k^2 M^{-1}||L_{K^{1/2}\varphi_k}||_K^2\\&\leq\sum_{k=\widetilde{M}+1}^{2\widetilde{M}} M^{-1}||L_{K^{1/2}\varphi_k}||_K^2\leq 1,\end{aligned}$$

which shows that $\mathcal{H}^*\subset\mathcal{F}$. The Varshamov-Gilbert bound shows that for any $M\geq 8$, there exists a set $\mathcal{B}=\{b^{(0)},b^{(1)},\ldots,b^{(N)}\}\subset\{0,1\}^M$such that

1. $b^{(0)}=(0,\ldots,0)'$;
2. $H(b,b')>M/8$ for any $b\neq b'\in\mathcal{B}$, where $H(\cdot,\cdot)=(1/4)\sum_{i=1}^M(b_i-b_i')^2$ is the Hamming distance;
3. $N\geq 2^{M/8}$.

The subset $\{\beta^{(0)},\ldots,\beta^{(N)}\}\subset\mathcal{F}^{q^*}$ is chosen as $\beta_j^{(i)}=\sum_{k=\widetilde{M}+1}^{2\widetilde{M}} b_{j,k-\widetilde{M}}^{(i)}M^{-1/2}$ $L_{K^{1/2}\varphi_k}$, $i=0,\ldots N$, $j=1,\ldots,q^*$. For any $0\leq i<j\leq N$, observe that

$$\|\beta^{(i)}-\beta^{(j)}\|_C^2=\sum_{l=1}^{q^*}\mathbb{E}_X\big(\langle\beta_l^{(i)}-\beta_l^{(i)},X\rangle\big)^2=\sum_{l=1}^{q^*}\sum_{k=\widetilde{M}+1}^{2\widetilde{M}}(b_{l,k-\widetilde{M}}^{(i)}-b_{l,k-\widetilde{M}}^{(j)})^2M^{-1}\rho_k.$$

Therefore,

$$\|\beta^{(i)}-\beta^{(j)}\|_C^2 \geq s_{2\widetilde{M}} M^{-1} \sum_{l=1}^{q^*}\sum_{k=1}^{\widetilde{M}} (b_{l,k}^{(i)}-b_{l,k}^{(j)})^2 \geq \frac{\rho_{2\widetilde{M}}}{2} \asymp \left(\frac{M}{q^*}\right)^{-2r},$$

and

$$\|\beta^{(i)}-\beta^{(j)}\|_C^2 \leq s_{\widetilde{M}} M^{-1} \sum_{l=1}^{q^*}\sum_{k=1}^{\widetilde{M}} (b_{l,k}^{(i)}-b_{l,k}^{(j)})^2 \leq \rho_{\widetilde{M}} \asymp \left(\frac{M}{q^*}\right)^{-2r}.$$

By taking M to be the smallest integer greater than $c_2\gamma^{-1/(2r+1)}(q^*)^{2r/(2r+1)}$ $n^{1/(2r+1)}$ with $c_2=(c_1\cdot 8\log 2)^{1/(1+2r)}$, the theorem is proved.

A3. Proof of Theorem 2

For any $q \geq q^*$, since $\hat{f}$ is the maximum, we have

$$-\ell(\hat{f})+\lambda\|\hat{\beta}\|_K^2 \leq -\ell(f^*)+\lambda\|\beta^*\|_K^2,$$

which gives

$$-\left(\ell(\hat{f})-\ell(f^*)\right)+\lambda\|\hat{\beta}\|_K^2 \leq \lambda\|\beta^*\|_K^2. \tag{A.1}$$

Define

$$d_h=\frac{\sqrt{h/h^*}-1}{\mathcal{H}(h,h^*)}.$$

Using $\log(1+x)\leq x$,

$$\begin{aligned}\ell(f)-\ell(f^*) &= \sum_{i=1}^{n} 2\log(1+\mathcal{H}(h,h^*)d_h(Y,X_i)) \\ &\leq \sum_{i=1}^{n} 2\mathcal{H}(h,h^*)d_h(Y,X_i) \\ &= 2\sqrt{n}\ \nu_n(d_h)\mathcal{H}(h,h^*)-n\ \mathcal{H}^2(h,h^*),\end{aligned}$$

where $\nu_n(g)=n^{-1/2}\sum_{i=1}^n(g(Y_i,X_i)-\mathbb{E}g(Y_i,X_i))$. So,

$$-(\ell(f)-\ell(f^*)) \geq n\ \mathcal{H}^2(h,h^*)-2\sqrt{n}\ \nu_n(d_h)\mathcal{H}(h,h^*).$$

Combining this with (A.1),

$$\begin{aligned}n\ \mathcal{H}^2(\hat{h},h^*)+\lambda\|\hat{\beta}\|_K^2 &\leq \lambda\|\beta^*\|_K^2+2\sqrt{n}\ \nu_n(d_{\hat{h}})\mathcal{H}(\hat{h},h^*) \\ &= \lambda\|\beta^*\|_K^2+2\sqrt{n}\ (\nu_n(g_{\hat{h}})-\nu_n(g_{h^*})),\end{aligned} \tag{A.2}$$

where $g_h = \sqrt{h/h^*} = \sqrt{f/f^*}$. It is critical to investigate the behavior of $|\nu_n(g_h) - \nu_n(g_{h^*})|$ as a function of $\mathcal{H}(h, h^*)$.

It follows from Lemma 1 that the bracketing entropy of $H_q(\epsilon)$ is

$$H_{[\,]}(\delta) = \log N_{[\,]}\Big(H_q(\epsilon), \delta\Big) \leq (10\big(C_2\delta^{-1/r} + 1\big)q + 1) \log\Big(\frac{C_1\ \epsilon}{\delta}\Big).$$

Then,

$$\int_0^\epsilon H_{[\,]}^{1/2}(\delta)d\delta \leq c\ \sqrt{q}\ \epsilon^{1-1/(2r)}.$$

The reminder of proof is identical to that in Section 5.6 of van de Geer (2001). We obtain

$$\sup_{f\in\mathcal{M}_q} \frac{\big|\nu_n(g_h) - \nu_n(g_{h^*})\big|}{\sqrt{q}\ \mathcal{H}(h, h^*)^{1-1/(2r)} \vee n^{-(r-1/2)/(2r+1)}} = O_p(1). \tag{A.3}$$

This allows us to conclude that

$$\begin{aligned}\sqrt{n}\ \big|\nu_n(g_{\hat{h}}) - \nu_n(g_{h^*})\big| &= \sqrt{n}\ \frac{\big|\nu_n(g_{\hat{h}}) - \nu_n(g_{h^*})\big|}{\sqrt{q}\ \mathcal{H}(\hat{h}, h^*)^{1-1/2r}}\sqrt{q}\mathcal{H}(\hat{h}, h^*)^{1-1/2r} \\ &= O_p(\sqrt{n})\ \sqrt{q}\ \mathcal{H}(\hat{h}, h^*)^{1-1/2r}.\end{aligned}$$

Combining this with (A.2) yields that $\mathcal{H}(\hat{h}, h^*) = O_p(n^{-r/(2r+1)})$ provided that λ is of order $n^{1/(2r+1)}$. This finishes the proof of Theorem.

A4. Proof of Theorem 3

First note that

$$\begin{aligned}&\overline{\lim_{n\to\infty}} \sup_{q>q^*} \frac{\Delta_n(q, q^*)}{\mathrm{pen}_n(q) - \mathrm{pen}_n(q^*)} \\ &\leq \lim_{n\to\infty} \sup_{q>q^*} \frac{q^{2r/(2r+1)}n^{1/(2r+1)}}{\mathrm{pen}_n(q) - \mathrm{pen}_n(q^*)} \overline{\lim_{n\to\infty}} \sup_{q>q^*} \frac{\Delta_n(q, q^*)}{q^{2r/(2r+1)}n^{1/(2r+1)}} = 0.\end{aligned}$$

Therefore, for all $q > q^*$.

$$\sup_{f\in\mathcal{M}_q} \Big(\ell_n(f) - \lambda_{q,n}\|\beta\|_K^2\Big) - \mathrm{pen}_n(q) < \sup_{f\in\mathcal{M}_{q^*}} \Big(\ell_n(f) - \lambda_{q^*,n}\|\beta\|_K^2\Big) - \mathrm{pen}_n(q^*).$$

This shows that $\overline{\lim}_{n\to\infty}\hat{q}_n \leq q^*$ a.s., which means that we do not asymptotically overestimate the order.

On the other hand, for any $q < q^*$,

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\Delta_n(q,q^*) \le \overline{\lim_{n\to\infty}} \sup_{f\in\mathcal{M}_q} \frac{1}{n}\sum_{j=1}^{n}\log\frac{f(Y_i|X_i)}{f^*(Y_i|X_i)} + \lim_{n\to\infty}\frac{\lambda_{q^*,n}}{n}\|\beta^*\|_K^2$$

which is strictly negative based on Part (b) of Lemma 2. Since $\text{pen}_n(q)/n \to 0$ as $n \to \infty$ for $q < q^*$, we have

$$\overline{\lim_{n\to\infty}}\frac{1}{n}\Big\{\Delta_n(q,q^*) - \text{pen}_n(q) + \text{pen}_n(q^*)\Big\} < 0, \quad a.s.$$

We obtain, for all $q < q^*$,

$$\sup_{f\in\mathcal{M}_q}\Big(\ell_n(f) - \lambda_{q,n}\|\beta\|_K^2\Big) - \text{pen}_n(q) < \sup_{f\in\mathcal{M}_{q^*}}\Big(\ell_n(f) - \lambda_{q^*,n}\|\beta\|_K^2\Big) - \text{pen}_n(q^*).$$

This shows that $\overline{\lim}_{n\to\infty}\hat{q}_n \ge q^*$ a.s., which means that we do not asymptotically underestimate the order.

References

Bowman, A. (2010). Functional Data Analysis with R and MATLAB. *Journal of Statistical Software, Book Reviews* **34**, 1–2.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics* **34**, 2159–2179.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *Journal of the American Statistical Association* **107**, 1201–1216.

Calinon, S., F., G. and Billard, A. (2007). On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **37**, 286–298.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics* **25**, 573–578.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics* **37**, 35–72.

Csiszár, I. and Shields, P. C. (2000). The consistency of the BIC Markov order estimator. *The Annals of Statistics* **28**, 1601–1619.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary arma processes. *The Annals of Statistics* **27**, 1178–1209.

Dong, A., Toledo, J. B., Honnorat, N., Doshi, J., Varol, E., Sotiras, A. et al. (2016). Heterogeneity of neuroanatomical patterns in prodromal Alzheimer's disease: Links to cognition, progression and biomarkers. *Brain* **140**, 735–747.

Du, P. and Wang, X. (2014). Penalized likelihood functional regression. *Statistica Sinica* **24**, 1017–1041.

Feng, X., Li, T., Song, X. and Zhu, H. (2020). Bayesian scalar on image regression with nonignorable nonresponse. *Journal of the American Statistical Association* **115**, 1574–1597.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis*. Springer Science, New York.

Fonov, V., Evans, A. C., Botteron, K., Almli, C. R., McKinstry, R. C., Collins, D. L. et al. (2011). Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage* **54**, 313–327.

Gassiat, E. and van Handel, R. (2012). Consistent order estimation and minimal penalties. *IEEE Transactions on Information Theory* **59**, 1115–1128.

Gassiat, E. and van Handel, R. (2014). The local geometry of finite mixtures. *Transactions of the American Mathematical Society* **366**, 1047 – 1072.

Ghahramani, Z. and Jordan, M. I. (1994). Supervised learning from incomplete data via an EM approach. *Advances in Neural Information Processing Systems* **6**, 120–127.

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence rates for functional linear regression. *The Annals of Statistics* **35**, 70–91.

Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics* **46**, 2844–2870.

Heneka, M. T., Carson, M. J., El Khoury, J., Landreth, G. E., Brosseron, F., Feinstein, D. L. et al. (2015). Neuroinflammation in Alzheimer's disease. *The Lancet Neurology* **14**, 388–405.

Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *The Annals of Statistics* **44**, 2726–2755.

Latta, C. H., Brothers, H. M. and Wilcock, D. M. (2015). Neuroinflammation in Alzheimer's disease; A source of heterogeneity and target for personalized therapy. *Neuroscience* **302**, 103–111.

Lin, F., Ren, P., Lo, R. Y., Chapman, B. P., Jacobs, A., Baran, T. M. et al. (2017). Insula and inferior frontal gyrus' activities protect memory performance against Alzheimer's disease pathology in old age. *Journal of Alzheimer's Disease* **55**, 669–678.

Liu, R. and Zhu, H. (2021). Statistical disease mapping for heterogeneous neuroimaging studies. *Canadian Journal of Statistics* **49**, 10–34.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Micchelli, C. and Wahba, G. (1981). Design problems for optimal surface interpolation. *Approximation Theory and Applications* (Edited by Z. Ziegler), 329–347. Academic Press, New York.

Nishii, R. (1988). Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis* **27**, 392–403.

Ombao, H., Lindquist, M., Thompson, W. and Aston, J. (2016). *Handbook of Neuroimaging Data Analysis*. CRC Press, Boca Raton.

Ramsay, J. and Sloverman, B. (2005). *Functional Data Analysis*. Springer Science, New York.

Richard, S. and Green, P. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society: Series B (Statistical Methodology)* **59**, 731–792.

Schölkopf, B. (2001). The kernel trick for distances. In *Advances in Neural Information Processing Systems 13* (Edited by T. K. Leen, T. G. Dietterich and V. Tresp), 301–307. MIT Press.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines*. Springer Science & Business Media, New York.

Stulp, F. and Sigaud, O. (2015). Many regression algorithms, one unified model: A review. *Neural Networks* **69**, 60–79.

Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*. Springer, New York.

van de Geer, S. A. (2000). *Applications of Empirical Process Theory*. Cambridge University Press, Cambridge.

van de Geer, S. A. (2001). *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge.

Wahba, G. (1990). *Spline Models for Observational Data*. Siam.

Wang, J. L., Chiou, J. M. and Muller, H. G. (2016). Review of functional data analysis. *Annual Review of Statistics and Its Application* **3**, 257–295.

Wang, X. and Ruppert, D. (2015). Optimal prediction in an additive functional model. *Statistica Sinica* **25**, 567–589.

Wang, X., Zhu, H. and ADNI (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association* **112**, 1156–1168.

Wang, Y. R., Li, L., Li, J. J. and Huang, H. (2021). Network modeling in biology: Statistical methods for gene and brain networks. *Statistical Science* **36**, 89–108.

Yao, F., Fu, Y. and Lee, T. (2011). Functional mixture regression. *Biostatistics* **12**, 341–353.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics* **38**, 3412–3444.

Zhu, H.-T. and Zhang, H. (2004). Hypothesis testing in mixture regression models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 3–16.

Xiao Wang

Department of Statistics, Purdue University, Indiana, West Lafayette, IN 47907, USA.

E-mail: wangxiao@purdue.edu

Leo Yu-Feng Liu

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, NC 27599-7400, USA.

E-mail: leo1986@unc.edu

Hongtu Zhu

Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599-7400, USA.

E-mail: htzhu@email.unc.edu

(Received May 2021; accepted November 2021)