Statistica Sinica **24** (2014), 1017-1041
doi:http://dx.doi.org/10.5705/ss.2012.235

PENALIZED LIKELIHOOD FUNCTIONAL REGRESSION

Pang Du and Xiao Wang

Virginia Tech and Purdue University

Abstract: This paper studies the generalized functional linear model with a scalar response and a functional predictor. The response given the functional predictor is assumed to come from the distribution of an exponential family. A penalized likelihood approach is proposed to estimate the unknown intercept and coefficient function in the model. Inference tools such as point-wise confidence intervals of the coefficient function and prediction intervals are derived. The minimax rate of convergence for the error in predicting the mean response is established. It is shown that the penalized likelihood estimator attains the optimal rate of convergence. Our simulations demonstrate a competitive performance against the existing approach. The method is further illustrated in the use of the DTI tractography to distinguish corpus callosum tracts with multiple sclerosis from normal tracts.

Key words and phrases: Confidence intervals, generalized functional regression, minimax rate of convergence, prediction error, penalized likelihood, reproducing kernel Hilbert space, .

1. Introduction

In the past decade, a lot of attention has focused on functional linear models, where the dependence of a continuous response variable on a functional covariate is modeled through a regression model, generalizing the standard multiple linear regression. The literature on functional linear models is vast and we only refer to the well-known monographs Ramsay and Silverman (2005) and Ferraty and Vieu (2006), and some recent developments such as Cai and Hall (2006), Hall and Horowitz (2007), Crambes, Kneip, and Sarda (2009), and Yuan and Cai (2010), for further references. Our paper is concerned with a general setting that includes functional linear models, functional logistic regression, functional log-linear regression, and functional censored regression as special cases.

Generalized functional regression models have not been as well studied. James and Hastie (2001) extended Fisher's linear discriminant analysis to handle the classification problem with functional covariates. James (2002) studied generalized linear models with functional predictors. These papers used natural cubic splines to approximate the functional covariate, but neither provided theoretical insights into the method and the choice of the number of basis functions is purely empirical. Müller and Stadtmüller (2005) considered functional

quasi-likelihood estimation of a generalized functional linear model and used a truncated orthonormal expansion to approximate the predictor process. The truncation parameter increases with the sample size in their asymptotic analysis and is selected by AIC in practice. The orthonormal basis functions often come from the leading functions in the Fourier basis system or the Karuhen-Loève expansion of the predictor process. Since the latter is used in functional principal component analysis (FPCA), this is often referred to as the FPCA approach. Some extensions of such an approach are Crainiceanu, Staicu, and Di (2009), where functional data are observed at multiple levels, and Li, Wang, and Carroll (2010), where the predictors may consist of scalar and functional covariates together with their interactions. Dou, Pollard, and Zhou (2010) use the FPCA approach to study the functional regression for general exponential families, and assume that the coefficient function is a linear combination of the eigenfunctions of the covariance kernel with decaying coefficients.

Despite its popularity the FPCA approach has a critical drawback. As shown in Cai and Yuan (2012), the success of the FPCA approach hinges crucially on the assumption that the coefficient function can be effectively represented by the leading functional principal components of the covariance kernel of the predictor process. This is problematic. The coefficient function and the covariance kernel of the predictor process are generally unrelated, so the FPCs from the covariance kernel may not form an efficient basis system for the coefficient function. For example, consider the covariance kernel $K(t,s) = \sum_{k=1}^{\infty}(|k-25|+1)^{-2}\psi_k(t)\psi_k(s)$ and the coefficient function $\beta(t) = \sum_{k=1}^{50} k^{-2}\psi_k(t)$, where $\{\psi_k : k = 1, 2, \ldots\}$ is an orthonormal basis function system. The leading eigen-functions of K are $\psi_k(t)$'s with k around 25, whereas an effective representation of β requires the leading functions in the system $\{\psi_k : k = 1, 2, \ldots\}$. These sets of functions do not overlap unless a large number of basis functions are used in the representation of β, but too many basis functions without any roughness penalty often lead to numerical instability and large variation.

The FPCA approach is a popular model regularization method for functional regression models. Here a finite-dimensional approximation to the functional parameter is used to regularize the model complexity, in theory and in practice. This approximation may result from a well-known basis system, such as the Fourier basis and the B-spline basis, or from the FPCs. This approach has a well-known drawback, as pointed out in Ramsay and Silverman (2005): the dimension of the approximation jumps in a discrete manner, yielding a discrete and often unprecise control on the model complexity, and can often lead to inaccurate functional estimate with hard-to-interpret "artificial" bumps. An alternative regularization approach uses a roughness penalty to govern model complexity via a smoothing parameter. This approach often yields more reasonable functional

estimates than its finite expansion competitor, and even leads to the advice "splines should be penalized", Carroll and Ruppert (2006).

The practical superiority of the roughness penalty approach often comes at the price of a harder theory. Crambes, Kneip, and Sarda (2009) and Yuan and Cai (2010) are among the few who have done thorough theoretical analysis of the roughness penalty approach to functional linear regression models. Cardot and Sarda (2005) is a first theoretical attempt in the direction of generalized functional regression models by penalized likelihood. They used penalized B-splines to estimate the functional parameter and derived the L^2 convergence rate of the estimation error. However, their convergence rate is not optimal possibly due to the choice of their spline basis, and they did not investigate the convergence rate of the error in predicting the mean response from the coefficient function estimator.

We adopt a penalized likelihood approach to study generalized functional regression models. The minimax rate of convergence for the error in predicting the mean response is established. It is shown that the penalized likelihood estimator attains this optimal rate of convergence. This is not a trivial extension of Yuan and Cai (2010), as they only considered the functional linear regression model, where the objective function is quadratic and a closed form solution is available for the estimate. We resort to the calculus of variation to explicitly characterize the penalized likelihood estimator, and, ultimately, to obtain the optimal rate of convergence. As demonstrated by Cai and Hall (2006) and Hall and Horowitz (2007) in functional linear regression models, the prediction error has different characteristics from the estimation error that measures the distance between the estimator and the true coefficient function. Particularly, the predicted mean response involves an integral of the coefficient function and is thus smoother. Hence, a procedure optimal in smoothing the coefficient function may not be optimal in predicting, or can actually over-smooth the prediction. Our method differs from that of Cardot and Sarda (2005) in that we use the smoothing spline basis functions that are naturally introduced by the derivative penalty of the coefficient function. Based on the Representer Theorem in Yuan and Cai (2010), this spline basis allows a finite-dimensional exact solution to the infinite-dimensional optimization problem of the corresponding penalized likelihood and hence greatly facilitates the estimation. We also derive confidence intervals for the coefficient function and the predicted mean response.

The rest of the paper is organized as follows. Section 2 proposes the penalized likelihood functional regression model. Section 3 introduces the Representer Theorem and the estimation procedure. Section 4 derives the point-wise confidence intervals for the coefficient function and the prediction intervals. Section 5 establishes the minimax convergence rate for the prediction error. Section 6

presents simulations comparing the proposed method with the competitor in Müller and Stadtmüller (2005). Section 7 applies the method to distinguishing corpus callosum tracts with multiple sclerosis from normal tracts based on their diffusivity profiles from the DTI tractography. Potential extensions and future research directions are discussed in Section 8. The technical proof of the main theorem is given in the Appendix.

2. The Model

Assume that the observed data (Y_i, X_i) are i.i.d. copies of (Y, X) for $i = 1, \ldots, n$, where $\{X(t) : t \in \mathcal{T}\}$ is a square integrable random function on a compact interval $\mathcal{T}$ with mean $\mu(t)$ and covariance $K(s,t) = \mathrm{cov}(X(s), X(t))$. Without loss of generality we take $\mathcal{T} = [0,1]$ and $\mu(t) = 0$. Given X, the response variable Y follows an exponential family distribution with density

$$\exp\Big[\frac{y\eta(X;\alpha_0,\beta_0) - b(\eta(X;\alpha_0,\beta_0))}{a(\phi)} + c(y,\phi)\Big], \tag{2.1}$$

where $\eta(X;\alpha_0,\beta_0) = \alpha_0 + \int_0^1 X(t)\beta_0(t)dt$ is the canonical parameter with an unknown constant α_0 and an unknown coefficient function $\beta_0(t)$, a, b, and c are known functions, and ϕ is either known or considered as a nuisance parameter that is independent of X. For example, a Gaussian response with variance σ^2 has $a(\phi) = \sigma^2$ and $b(\eta) = \eta^2/2$, a Bernoulli response has $a(\phi) = 1$ and $b(\eta) = \log(1 + e^\eta)$, and a Poisson response has $a(\phi) = 1$ and $b(\eta) = e^\eta$. For notational simplicity, we choose not to include α_0 into the notation of η with the dependence implicitly assumed. Also note that (2.1) can be easily generalized to the quasi-likelihood estimation setting as in Müller and Stadtmüller (2005).

The infinite-dimensional function $\beta_0(t)$ is not estimable by n observations without additional constraints on the functional regression model. Müller and Stadtmüller (2005) dealt with this by modeling $\beta_0(t)$ as a truncated expansion of certain orthonormal basis functions. They allow the number of basis functions included in the truncation to increase with n for asymptotic analysis. We consider a roughness penalty approach. Let $\mathcal{L}_2 \equiv \mathcal{L}_2([0,1]) = \{f : \int_0^1 f^2 dt < \infty\}$ and assume that β_0 belongs to the Sobolev space $W_2^m([0,1])$ of order m,

$$W_2^m([0,1]) = \{\beta : [0,1] \to \mathbb{R} | \beta, \beta', \ldots, \beta^{(m-1)} \text{ are absolutely continuous and } \beta^{(m)} \in \mathcal{L}_2\}.$$

We propose to find estimates $\hat{\alpha} \in \mathbb{R}$ of α_0 and $\hat{\beta} \in W_2^m$ of β_0 that minimize the penalized likelihood

$$-\frac{1}{n}\sum_{i=1}^n \Big\{Y_i\eta(X_i;\beta) - b(\eta(X_i;\beta))\Big\} + \frac{\lambda}{2}\int_0^1 [\beta^{(m)}(t)]^2 dt, \tag{2.2}$$

where the sum is the negative log likelihood up to a constant derived from the density (2.1) representing the goodness-of-fit of the estimate, $\int_0^1[\beta^{(m)}(t)]^2dt$ is the roughness penalty, and $\lambda > 0$ is the smoothing parameter balancing the tradeoff. The terms $c(Y_i, \phi)$ are independent of η and thus are dropped from (2.2), and the dispersion parameter $a(\phi)$ is absorbed into λ.

3. Estimation

In this section, we introduce the Representer Theorem of Yuan and Cai (2010) which guarantees a finite-dimensional representation of the estimate $\hat{\beta} \in W_2^m([0,1])$.

Consider the inner product

$$\langle f, g\rangle = \sum_{d=0}^{m-1}\int_0^1 f^{(d)}\int_0^1 g^{(d)} + \int_0^1 (f^{(m)}g^{(m)}) \tag{3.1}$$

that makes W_2^m a *reproducing product Hilbert space*. The roughness penalty $J(\beta) := \int_0^1[\beta^{(m)}(t)]^2dt$ in (2.2) defines a squared semi-norm on W_2^m. The null space of J, $\mathcal{H}_0 := \{\beta \in W_2^m : J(\beta) = 0\}$, is an m-dimension linear subspace of W_2^m with an orthonormal basis $\{k_0(t), k_1(t), \ldots, k_{m-1}(t)\}$, where $k_0(t) = 1$ and, for $r \geq 1$, $k_r(t) = B_r(t)/r!$ is the rth scaled Bernoulli polynomial. Denote by $\mathcal{H}_1$ the orthogonal complement of $\mathcal{H}_0$ in W_2^m such that $W_2^m = \mathcal{H}_0 \oplus \mathcal{H}_1$. Then $\mathcal{H}_1$ forms a reproducing kernel Hilbert space with the inner product (3.1) restricted to $\mathcal{H}_1$. Particularly, the reproducing kernel of $\mathcal{H}_1$ has the form

$$R(s,t) = k_m(s)k_m(t) + (-1)^{m-1}k_{2m}(|s-t|). \tag{3.2}$$

We refer to Wahba (1990) and Gu (2002) for more details on reproducing kernel Hilbert spaces.

The reproducing kernel function R defines a nonnegative definite operator from $\mathcal{L}_2$ to $\mathcal{H}_1$, $(Rf)(t) = \int_0^1 R(t,s)f(s)ds \in \mathcal{H}_1$ for any $f \in \mathcal{L}_2$. The Representer Theorem in Yuan and Cai (2010) states that the minimizer $\hat{\beta}$ of the penalized likelihood (2.2) in $W_2^m([0,1])$ lies in the subspace of functions of the form

$$\beta(t) = \sum_{j=0}^{m-1} d_j k_j(t) + \sum_{i=1}^{n} c_i(RX_i)(t) \equiv \boldsymbol{\phi}'(t)\mathbf{d} + \boldsymbol{\xi}'(t)\mathbf{c}, \tag{3.3}$$

where $\mathbf{d} = (d_0, \ldots, d_{m-1})'$ and $\mathbf{c} = (c_1, \ldots, c_n)'$ are vectors of unknown coefficients, and $\boldsymbol{\phi}' = (k_0, \ldots, k_{m-1})$ and $\boldsymbol{\xi}' = (RX_1, \ldots, RX_n)$ are vectors of basis functions.

The objective function (2.2) is strictly convex in η, which is a linear transformation of β. Hence for a fixed smoothing parameter λ, we can use the

Newton-Raphson procedure to compute the minimizer $(\hat{\alpha}_\lambda, \hat{\beta}_\lambda)$ of (2.2). Given an estimate $(\tilde{\alpha}, \tilde{\beta})$, define $\tilde{w}_i = b''(\eta(X_i; \tilde{\beta}))$ and $\tilde{Y}_i = \eta(X_i; \tilde{\beta}) - \{b'(\eta(X_i; \tilde{\beta})) - Y_i\}/b''(\eta(X_i; \tilde{\beta}))$. Then the quadratic approximation of $-Y_i\eta(X_i; \beta) + b(\eta(X_i; \beta))$ at $(\tilde{\alpha}, \tilde{\beta})$ is $\frac{1}{2}\tilde{w}_i\{\tilde{Y}_i - \eta(X_i; \beta)\}^2 + C_i$, where C_i is independent of (α, β). The Newton iteration updates $(\tilde{\alpha}, \tilde{\beta})$ by the minimizer of the penalized weighted least squares functional

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i\{\tilde{Y}_i - \eta(X_i; \beta)\}^2 + \frac{\lambda}{2}\int_0^1 [\beta^{(m)}(t)]^2 dt. \tag{3.4}$$

Plugging in the expression (3.3), we have $\eta(X_i; \beta) = \alpha + (\int_0^1 \boldsymbol{\phi}(t)X_i(t)dt)'\mathbf{d} + (\int_0^1 \boldsymbol{\xi}(t)X_i(t)dt)'\mathbf{c}$. Note that $\int_0^1 [\beta^{(m)}(t)]^2 dt = \mathbf{c}'\Sigma\mathbf{c}$, where Σ is an $n \times n$ matrix with $\Sigma_{ij} = \int_0^1 \int_0^1 X_i(s)R(t, s)X_j(t)dsdt$. With $\mathbf{d}^+ = (\alpha, \mathbf{d}')'$, (3.4) is quadratic in $\mathbf{c}$ and $\mathbf{d}^+$ whose closed form solution is easily available.

To select the smoothing parameter λ, we use the generalized approximate cross-validation (GACV) score developed in Xiang and Wahba (1996). It provides a cross-validation approximation to the Kullback-Leibler distance between an estimate $(\hat{\alpha}_\lambda, \hat{\beta}_\lambda)$ and the true (α_0, β_0). Its explicit form is

$$\begin{aligned}\mathrm{GACV}(\lambda) = &-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\eta(X_i; \hat{\beta}_\lambda) - b(\eta(X_i; \hat{\beta}_\lambda))\} \\ &+\frac{\mathrm{trace}(H)}{n - \mathrm{trace}(HW)}\frac{1}{n}\sum_{i=1}^{n} Y_i(Y_i - b'(\eta(X_i; \hat{\beta}_\lambda))),\end{aligned} \tag{3.5}$$

where $W = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_n)$ and $H = (Z'Z + n\lambda\Sigma)^{-1}Z'\tilde{\mathbf{Y}}$ is the hat matrix for solving (3.4). Here Z is the design matrix with all 1's in the first column, followed by the columns $\int_0^1 \phi_j(t)\mathbf{X}(t)dt, j = 0, \ldots, m-1$, and $\int_0^1 \xi_i(t)\mathbf{X}(t)dt, i = 1, \ldots, n$.

4. Confidence Intervals

In this section we derive confidence interval estimates for β_0 and the predicted mean response $E(Y|X_{n+1})$ with $X_{n+1}(\cdot)$ a new random covariate function with the distribution of $X_i, i = 1, \ldots, n$. Both are based on the confidence intervals for the coefficient vectors $\mathbf{c}$ and $\mathbf{d}^+$ that are derived from the Baysian model interpretation of the penalized likelihood (2.2). This generalizes a similar idea in nonparametric regression by smoothing splines, Wahba (1990).

With the decomposition $W_2^m = \mathcal{H}_0 \oplus \mathcal{H}_1$, $\beta = \beta_{\langle 0\rangle} + \beta_{\langle 1\rangle}$ with $\beta_{\langle 0\rangle} \in \mathcal{H}_0$ and $\beta_{\langle 1\rangle} \in \mathcal{H}_1$. Suppose we assign a diffuse prior to α in $\mathbb{R}$ and $\beta_{\langle 0\rangle}$ in $\mathcal{H}_0$. With $\beta_{\langle 0\rangle} = \boldsymbol{\phi}'\mathbf{d}$, we have a diffuse prior on $\mathbf{d}^+$. Similarly, we can assign to $\beta_{\langle 1\rangle}$ a mean 0 Gaussian process with a chosen covariance such that it yields the Gaussian prior $N(\mathbf{0}, \Sigma^{-1}/(n\lambda))$ for $\mathbf{c}$. Then the coefficient vector $\hat{\mathbf{b}} = (\hat{\mathbf{d}}^{+\prime}, \hat{\mathbf{c}}')'$

of the minimizer of the penalized likelihood (2.2) is be the posterior mode under such priors.

Through a second-order Taylor expansion of the log likelihood at this posterior mode, we can approximate the posterior distribution of $\mathbf{b} = (\mathbf{d}^{+'}, \mathbf{c}')'$ by a Gaussian distribution with mean $\hat{\mathbf{b}}$ and covariance H^{-1}/n. Hence the approximate posterior mean of $\beta(t)$ is $\hat{\beta}(t)$ and the approximate posterior variance is $s^2(t) = (0, \boldsymbol{\psi}'(t))H^{-1}\binom{0}{\boldsymbol{\psi}(t)}/n$, where $\boldsymbol{\psi}(t) = (\boldsymbol{\phi}'(t), \boldsymbol{\xi}'(t))'$. This yields a $100(1-\alpha)\%$ confidence interval $\hat{\beta}(t) \pm z_{\alpha/2}s(t)$ for β_0 at a point t. Similarly, a $100(1-\alpha)\%$ confidence interval for the mean response $E(Y|X_{n+1})$ can be obtained as

$$b'\Big(\eta(X_{n+1};\hat{\beta}) \pm z_{\alpha/2}\sqrt{(1, \int_0^1 \boldsymbol{\psi}'(t)X_{n+1}(t)dt)H^{-1}(1, \int_0^1 \boldsymbol{\psi}'(t)X_{n+1}(t)dt)'/n} \quad \Big)$$

provided b' is strictly monotone, as in the binomial and Poisson regression models.

5. Minimax Rate of Convergence

We have $E(Y_i|X_i) = b'(\eta(X_i;\beta_0))$ and $\mathrm{Var}(Y_i|X_i) \propto b''(\eta(X_i;\beta_0))$. As pointed out by Cai and Hall (2006), in terms of convergence rates the problems of estimating $\alpha + \int_0^1 X(t)\beta(t)dt$ and $\int_0^1 X(t)\beta(t)dt$ are not intrinsically different. Let $\alpha = 0$ in this discussion. Our goal is to derive the optimal rate of convergence of the error in predicting the conditional mean.

The conditional mean of Y_{n+1} for any new random function X_{n+1} possessing the same distribution as X and independent of $X_1, \ldots, X_n$ is $\eta(X_{n+1};\beta_0)$. The goal of prediction is to recover $\eta(X_{n+1};\beta_0)$ from the training data (X_i, Y_i), $i = 1, \ldots, n$. We measure its accuracy by the risk $E\Big\{\big[b'(\eta(X_{n+1};\hat{\beta})) - b'(\eta(X_{n+1};\beta_0))\big]^2 \Big| \hat{\beta}\Big\}$. If b'' is uniformly bounded, this risk is bounded by $\|\hat{\beta} - \beta_0\|_K^2$, where $K(t,s) = \mathrm{cov}(X(t), X(s))$ and

$$\|\hat{\beta} - \beta_0\|_M^2 = \int\int (\hat{\beta}(s) - \beta_0(s))M(s,t)(\hat{\beta}(t) - \beta_0(t))dsdt,$$

for any positive definite function M. We require some assumptions.

A1. β_0 belongs to the Sobolev space $W_2^m([0,1])$ of order m.

A2. The function b' is monotonic with $c_1 \le \inf_t b''(t) \le \sup_t b''(t) \le c_2$ and $\sup_t |b^{(3)}(t)| \le c_3$ for three positive constants c_1, c_2, c_3. The function $a(\cdot)$ has a positive lower bound.

A3. The covariance kernel K satisfies the Sacks-Ylvisaker conditions of order $r-1$, r a positive integer.

A4. For any function $f \in L_2[0,1]$, there exists a constant $L > 0$ such that

$$E\Big(\int_0^1 X(t)f(t)dt\Big)^4 \le L\left(E\Big(\int_0^1 X(t)f(t)dt\Big)^2\right)^2. \tag{5.1}$$

The detailed discussions of the Sacks-Ylvisaker conditions are in Sacks and Ylvisaker (1966, 1968, 1970) and in Ritter et al. (1995). A covariance kernel $K(t,s)$ satisfies the Sacks-Ylvisaker conditions of order zero is it is twice differentiable when $t \neq s$ but not differentiable when $t = s$. Then, K satisfies the Sacks-Ylvisake conditions of order $r-1$ if $\partial^{2(r-1)}K/\partial t^{r-1}\partial s^{r-1}$ satisfies the Sacks-Ylvisaker conditions of order zero. It follows from A3 that the eigenvalues ν_k of K satisfy $\nu_k \asymp k^{-2r}$. The process X has exactly $r-1$ derivatives in the mean squared sense, and $X^{(r-1)}$ is Lipschitz in the mean square sense. It follows from (5.1) that the linear functions of X have bounded kurtosis. This is true particularly when X is a Gaussian process with $c_4 = 3$. Let $\mathcal{F}(r,L)$ be the collection of the distributions F of the process X that satisfy A3 and A4.

We have the following result.

Theorem 1. *Let c_n be any sequence of positive numbers with $c_n \to 0$ as $n \to \infty$. Under* A1−A4*,*

$$\lim_{n\to\infty} \inf_{\hat\beta} \sup_{F\in\mathcal{F}(r,L),\beta_0\in W_2^m} P\Big(\|\hat\beta - \beta_0\|_K^2 \ge c_n n^{-2(m+r)/[2(m+r)+1]}\Big) = 1, \tag{5.2}$$

where the infimum is taken over all possible estimators $\hat\beta$ based on the training data (Y_i, X_i), $i = 1,\dots,n$.

The next result has the penalized smoothing spline estimator with the optimal rate of convergence.

Theorem 2. *Let $\hat\beta$ minimize the penalized likelihood* (2.2)*. If* A1−A4 *hold, then as $n \to \infty$,*

$$\|\hat\beta - \beta_0\|_K^2 = O_p\Big(\lambda + \frac{1}{n}\lambda^{-1/[2(m+r)]} + n^{-1}\Big). \tag{5.3}$$

From Theorem 2, the optimal choice of λ is of order $n^{-2(m+r)/[2(m+r)+1]}$, and thus the convergence rate of the error in predicting the conditional mean is of order $n^{-2(m+r)/[2(m+r)+1]}$. From Theorem 1 this is the optimal rate.

6. Numerical Studies

We present some simulation results comparing our method with that of Müller and Stadtmüller (2005), hereafter MS. For the MS method, we used its up-to-date implementation in the PACE package publicly available from the website of the Department of Statistics at U.C. Davis. PACE uses the leading functional

principal components as the orthonormal basis functions. For the predictor input to PACE, we used a vector of 200 values at a dense grid of points to represent each predictor curve. We conducted the simulations on both binary and Poisson data. Due to the similarity of the results, we only present the results on binary data.

We considered three simulation settings. For simplicity, the intercept α was set to zero in all the settings.

- Setting 1: This matches well with Müller and Stadtmüller (2005). Let $\varphi_k(t) = \sqrt{2}\sin(k\pi t)$ for $k \geq 1$. The random predictor function X was generated as $X = \sum_{k=1}^{20} \epsilon_k \varphi_k(t)$ with $\epsilon_k \sim N(0, 1/k^2)$. The true coefficient function was $\beta_0(t) = \varphi_1(t) + \varphi_2(t)/2 + \varphi_3(t)/3$.
- Setting 2: This was adopted from Yuan and Cai (2010). The true coefficient function was $\beta_0(t) = \sum_{k=1}^{50} 4(-1)^{k+1}k^{-2}\phi_k(t)$, where $\phi_1(t) = 1$ and $\phi_{k+1}(t) = \sqrt{2}\cos(k\pi t)$ for $k \geq 1$. The random predictor function X was generated as $X = \sum_{k=1}^{50} \zeta_k Z_k \phi_k$, with Z_k independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ and $\zeta_k = (-1)^{k+1}k^{-1}$.
- Setting 3: Here $\beta_0(t) = \sum_{k=1}^{50} 4k^{-2}\phi_k(t)$ and $X(t) = \sum_{k=1}^{50}(|k-5|+1)^{-2}Z_k\phi_k$, where Z_k were the same as in Setting 2. As discussed, the MS method may have difficulty with this setting due to the disordered representative basis functions for $X(t)$ and $\beta_0(t)$.

For each setting, we generated 100 data replicates with sample size $n = 500$, as well as an independent testing data set with sample size 10,000 to evaluate the prediction performance. For each training data set, we computed the estimates of $\beta_0(t)$ and the 95% point-wise confidence intervals of $\beta_0(t)$ on a fine grid of t based on the formula in Section 4. For each estimate of $\beta_0(t)$ thus obtained, we computed the predicted mean and the 95% mean prediction intervals for each observation in the testing data. The following performance metrics were computed for our method and the MS method: estimation error defined as the integrated squared error of the estimate, prediction error defined as the mean squared errors of predicting η on the testing data, and misclassification error which is the rate of misclassification on the testing data. To assess the effect of sample size, we also repeated simulation Setting 2 for sample sizes $n = 250$ and 1,000.

Figure 1 shows the mean estimates of $\beta_0(t)$ in the middle of the plots compared with the true functions. In Setting 1, the estimates from our method and the MS method match almost exactly with the true function except for a small deviation at the left end for our estimate. In Setting 2, the estimates also track the true function well; they are close to each other except at the ends of the interval where information from data dwindles. In Setting 3, our method

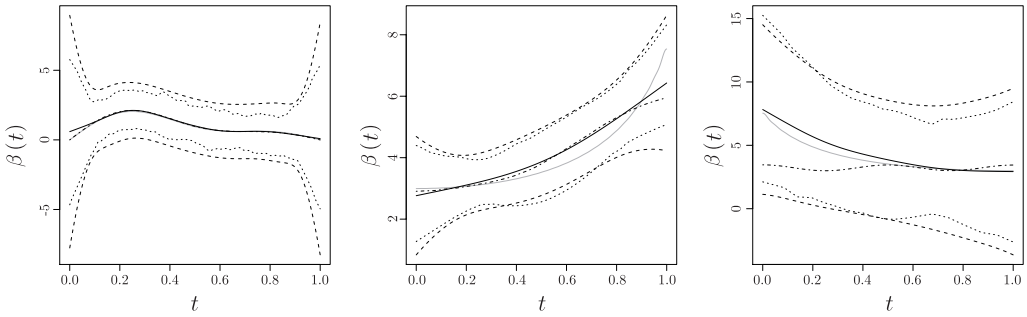

Figure 1. Estimates of $\beta_0(t)$ based on 100 data replicates: Setting 1 (left), Setting 2 (middle), and Setting 3 (right). Faded solid: true function; solid: our mean estimates; dashed: connected means of point-wise 95% confidence limits; dotted: empirical 2.5th and 97.5th percentiles from the 100 function estimates; dot-dashed: mean estimates from the method of Müller and Stadtmüller (2005).

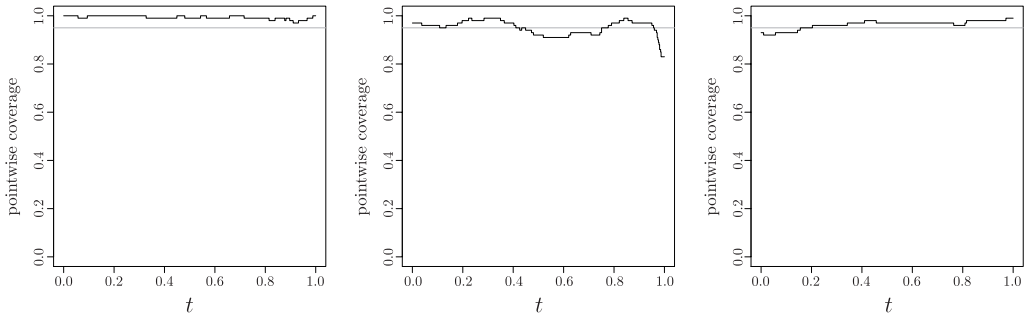

Figure 2. Point-wise empirical coverage of 95% confidence intervals of $\beta_0(t)$ based on 100 data replicates: Setting 1 (left), Setting 2 (middle), and Setting 3 (right). Faded line: nominal level 0.95; stepped line: connected point-wise coverage.

still yields an accurate estimate but the MS method suffers, as expected. This is because the true coefficient function cannot be effectively represented by the leading eigen-functions of the true covariance kernel function.

The performance metrics in Table 1 show favorable results for our method in the last two simulation settings. Particularly for Setting 3, the large estimation error of the MS method is no surprise, considering the deviation we see in Figure 1 at the first half of the interval. Even under Setting 1, designed to match the assumptions of the MS method, our method can yield competitive performance metrics.

Figure 1 shows the connected means of the 100 point-wise 95% confidence limits at the grid points, against the connected empirical 2.5th and 97.5th percentiles of the 100 function estimates at the grid points. Figure 2 shows the

Table 1. Performance metrics of the proposed method and the method of Müller and Stadtmüller (2005) for three simulation settings with $n = 500$. The numbers are means and standard deviations (in brackets). The mean of estimation error was computed over 100 data replicates. The mean of prediction error and misclassification error were computed over a testing data set with 10,000 new observations.

	Method	Estimation Error	Prediction Error	Misclassification Error
Setting 1	Ours	0.21(0.23)	0.051(0.031)	0.106(0.003)
	MS	0.17(0.15)	0.046(0.030)	0.105(0.003)
Setting 2	Ours	0.53(0.80)	0.27(0.66)	0.100(0.001)
	MS	0.83(1.08)	0.34(0.79)	0.111(0.090)
Setting 3	Ours	0.68(0.57)	0.027(0.018)	0.360(0.006)
	MS	7.30(0.13)	0.053(0.015)	0.371(0.015)

Table 2. Performance metrics of the proposed method and the method of Müller and Stadtmüller (2005) for Setting 2 with $n = 250$ and 1,000. The numbers are means and standard deviations (in brackets). The mean of estimation error was computed over 100 data replicates. The mean of prediction error and misclassification error were computed over a testing data set with 10,000 new observations.

	Method	Estimation Error	Prediction Error	Misclassification Error
$n = 250$	Ours	0.85(1.19)	0.46(1.05)	0.101(0.002)
	MS	1.49(1.53)	0.59(0.95)	0.113(0.090)
$n = 1000$	Ours	0.29(0.25)	0.12(0.17)	0.100(0.001)
	MS	0.44(0.37)	0.14(0.17)	0.110(0.090)

point-wise empirical coverage of the confidence intervals for the three settings, calculated as the percentage of times the intervals computed from a data replicate covers the true function. Under Setting 1, we see a conservative performance of the intervals; under Settings 2 and 3, the intervals delivered a good performance. Figure 1 shows that the confidence intervals and empirical quantiles, are very close to each other. Most of the empirical coverage values in Figure 2 are also close to the nominal level. The empirical coverages of the 95% prediction intervals calculated on the testing data were, respectively, 0.971, 0.947 and 0.968 for the three simulations settings.

7. Application: DTI Tractography

In this section we focus on an application to classifying corpus callosum tracts as from multiple sclerosis patients or control patients. A similar study on classifying white matter tracts was presented in Goldsmith et al. (2010). The corpus callosum is the largest white matter structure in the brain connecting the left and right cerebral hemispheres. The tracts within it consist of axons sur-

rounded by fatty myelin sheaths that play a critical role in signal transmission. Multiple sclerosis is a brain disease in which the myelin sheaths around the axons are damaged, leading to severe disability in an affected individual. The diffusion tensor imaging (DTI) tractography is a magnetic resonance imaging (MRI) technique that allows the study of white-matter tracts by measuring the diffusivity of water in them. Hence, the functional predictor used here for classifying a tract is the diffusivity profile of the tract. The data set is part of the R `refund` package and consists of 334 cases and 42 controls. The diffusivity profiles for a random sample of three case tracts and three control tracts are plotted in the left panel of Figure 3. This data set, including the functional predictor considered, is different from the one in Goldsmith et al. (2010) although we consider a similar classification problem.

The fitted coefficient functions for our method and that of Müller and Stadtmüller (2005) are displayed in the right panel of Figure 3 with the 95% confidence limits. Both estimates show a similar overall trend: the tracts whose diffusivity profile is above average between distances 80 and 90 are more likely to be multiple sclerosis cases; the tracts whose diffusivity profile is above average between distances 55 and 80 are more likely to be controls; the diffusivity profile between distances 20 and 55, the flatter regions of the diffusivity profiles of the six sample tracts in Figure 3, does not seem to be as important as the other parts of the tract. The MS estimate is a bit jumpier on the ends of the interval, a commonly-seen phenomenon when a basis expansion is used without penalty. The delete-one misclassification rates were 10.6% for our method and 11.4% for the MS method.

8. Discussion

We have considered the penalized likelihood approach for generalized functional linear regression with the assumption that the functional covariate $X(t)$ is completely observed. When $X(t)$ is observed only at discrete times, in particular, when the time points are sparsely sampled, a popular strategy to handle these functional data is a two-stage procedure. At the first step, a nonparametric method is used for data from each curve to obtain an estimate of X. The interesting readers are referred to Hall and Hosseini-Nasab (2006) for discussion of two-step procedures. The technique proposed here can be applied to sparsely sampled data directly without using two steps. Its development is involved and will be reported in the future.

We have provided an optimal rate estimator of predicting the mean response. Interesting questions, such as the asymptotic distribution of the estimator and adaptive estimation over the Sobolev classes, remain open. A recent work in this

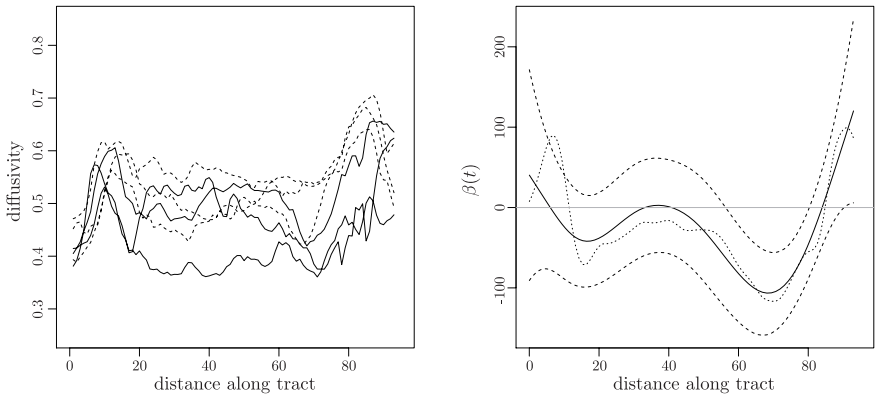

Figure 3. Left: Diffusivity profiles for a random sample of three case tracts (solid) and three control tracts (dashed). Right: Coefficient function estimates for the DTI tractography example, the coefficient estimate (solid) and the connected point-wise 95% confidence limits (dashed) from our method, and the estimate from the method of Müller and Stadtmüller (2005) (dotted). The faded solid line is the zero reference line.

direction is Cai and Yuan (2012) where adaptive estimation for the Gaussian case is studied.

Acknowledgements

The authors thank the Editor, the Associate Editor and two referees for their insightful comments that have significantly improved the paper. Du and Wang's research has been supported by the NSF grants DMS-1007126, DMS-1042967, and CMMI-0900753.

Appendices

A.1. Proof of Theorem 1

First, any lower bound for a specific case yields immediately a lower bound for the general case. We only consider the case when $\alpha_0 = 0$. Denote by J_m the reproducing kernel for W_2^m, $W_2^m = \mathcal{H}(J_m)$. Let $(\rho_k, \psi_k)_{k\geq 1}$ be the eigenvalue and eigenfunction pairs for $J_m^{1/2} K J_m^{1/2}$. Following (Cai and Yuan, 2012, Theorem 5), we can show that $\rho_k \asymp k^{-2(m+r)}$ under A3. Let M be the smallest integer which is larger than $c_0 n^{-1/[2(m+r)+1]}$ and we specify c_0 later. For any $\tau = (\tau_{M+1}, \ldots, \tau_{2M})^T \in \{0,1\}^M$, let

$$\beta_\tau(t) = \sum_{k=M+1}^{2M} \tau_k M^{-1/2} J_m^{1/2} \psi_k(t).$$

Observe that $\beta_\tau \in W_2^m$, since

$$\left\|\beta_\tau^{(m)}\right\|_2^2 = \left\|\beta_\tau\right\|_{\mathcal{H}(J_m)}^2 = \Big\| \sum_{k=M+1}^{2M} \tau_k^2 M^{-1} J_m^{1/2}\psi_k(t)\Big\|_{\mathcal{H}(J_m)}^2$$
$$= \sum_{k=M+1}^{2M} \tau_k^2 M^{-1}\left\|J_m^{1/2}\psi_k(t)\right\|_{\mathcal{H}(J_m)}^2 \le \sum_{k=M+1}^{2M} M^{-1}\left\|J_m^{1/2}\psi_k(t)\right\|_{\mathcal{H}(J_m)}^2 = 1.$$

Here we use the fact that

$$\left\langle J_m^{1/2}\psi_j, J_m^{1/2}\psi_k\right\rangle_{\mathcal{H}(J_m)} = \left\langle \psi_j, J_m\psi_k\right\rangle_{\mathcal{H}(J_m)} = \left\langle \psi_j, \psi_k\right\rangle_{L_2} = \delta_{jk}.$$

We apply the results from Tsybakov (2009) to establish the minimax lower bound based on testing multiple hypotheses. Fix $\alpha \in (0, 1/8)$. We have to check three conditions:

(a) $\beta_{\tau^j} \in W_2^m$, $j = 1, \ldots, N$;
(b) $\|\beta_{\tau^k} - \beta_{\tau^j}\|_K^2 \ge 2s > 0$, $\forall k \ne j$, and we specify s later;
(c) $1/N \sum_{j=1}^N \mathcal{K}(P_{\beta_{\tau^j}}, P_{\beta_{\tau^0}}) \le \alpha \log N$, where $\mathcal{K}(\cdot,\cdot)$ is the Kullback-Leibler distance between P_{β_τ} and $P_{\beta_{\tau'}}$, and P_β is the joint distribution of $\{(Y_i, X_i) : i = 1, \ldots, n\}$ with the coefficient function β.

If these conditions are satisfied, the minimax lower bound has the same order as s. First consider (a). For each τ_j, we have shown that $\beta_{\tau^j} \in W_2^m$. Further, the Varshamov-Gilbert bound shows that for any $M \ge 8$, there exists a set $\mathcal{B} = \{\tau^0, \tau^1, \tau^2, \ldots, \tau^N\} \subset \{0,1\}^M$ such that $\tau^0 = (0, \ldots, 0)^T$, $H(\tau, \tau') > M/8$ for any $\tau \ne \tau' \in \mathcal{B}$, where $H(\cdot,\cdot)$ is the Hamming distance, and $N > 2^{M/8}$. To check (b), we have

$$\|\beta_\tau - \beta_{\tau'}\|_K^2 = \sum_{k=M+1}^{2M} (\tau_k - \tau_k')^2 \rho_j M^{-1}$$
$$\ge cM^{-1} \sum_{k=M+1}^{2M} (\tau_k - \tau_k')^2 k^{-2(m+r)}$$
$$\ge cM^{-1}(2M)^{-2(m+r)}\frac{M}{8} = c2^{-2(m+r)-3}M^{-2(m+r)}.$$

For (c), observe that for any $\beta_\tau, \beta_{\tau'}$ we have

$$\log(P_{\beta_{\tau'}}/P_{\beta_\tau}) = \sum_{i=1}^n \frac{1}{a(\phi)}\Big[Y_i\eta(X_i;\beta_{\tau-\tau'}) - \big(b(\eta(X_i;\beta_\tau)) - b(\eta(X_i;\beta_{\tau'}))\big)\Big].$$

Therefore, by A1, for some positive constant $\tilde{c}_1$,

$$
\begin{aligned}
&\mathcal{K}(P_{\beta_{\tau'}}, P_{\beta_\tau}) \\
&\quad = \frac{n}{a(\phi)} E_X \Big[b'(\eta(X;\beta_{\tau'}))b(\eta(X;\beta_{\tau-\tau'})) - \big(b(\eta(X;\beta_\tau)) - b(\eta(X;\beta_{\tau'}))\big)\Big] \\
&\quad = \frac{n}{a(\phi)} E_X \Big[b'(\eta(X;\beta_{\tau'}))b(\eta(X;\beta_{\tau-\tau'})) - b'(\eta^*)b(\eta(X;\beta_{\tau-\tau'}))\Big] \\
&\quad \le \tilde{c}_1 n \|\beta_\tau - \beta_{\tau'}\|_K^2 = \tilde{c}_1 n \sum_{k=M+1}^{2M} (\tau_k - \tau_k')^2 j^{-2m} M^{-1} \rho_j \\
&\quad \le \tilde{c}_2 n M^{-2(m+r)}.
\end{aligned}
$$

Then for any $0 < \alpha < 1/8$,

$$
\frac{1}{N}\sum_{k=1}^{N} \mathcal{K}(P_{\beta_{\tau^j}}, P_{\beta_{\tau^0}}) \le \tilde{c}_2 n M^{-2(m+r)} \le \alpha \log 2^{M/8} \le \alpha \log N
$$

by taking $c_0 = c\alpha^{-1/[2(m+r)+1]}$ with a large numerical constant $c > 0$. Note that s in (b) is of order $n^{-2(m+r)/[2(m+r)+1]}$. This completes the proof of the theorem.

A.2. Proof of Theorem 2

We give a sketch of the proof with the proofs of all technical lemmas collected in the online Supplementary Material.

Write

$$
\epsilon_i = \frac{1}{\sqrt{b''(\eta(X_i;\beta_0))}}\Big(Y_i - E(Y_i|X_i)\Big), \qquad i = 1, \ldots, n.
$$

The ϵ_i are independent with mean zero and constant variance σ^2. Since $E(\epsilon_i|X_i) = E(\epsilon_i X_i) = 0$, $i = 1, \ldots, n$,

$$
Y_i = b'(\eta(X_i;\beta_0)) + \sqrt{b''(\eta(X_i;\beta_0))}\ \epsilon_i, \quad i = 1, \ldots, n.
$$

We first give the optimality conditions for $\hat{\beta}$. For any process X, let $X_i^{(-1)}(t) = \int_0^t X_i(s)ds$ and, for $k \ge 2$, $X_i^{(-k)}(t) = \int_0^t X_i^{(-k+1)}(s)ds$.

Lemma A.1. The necessary and sufficient conditions for $\hat{\beta}$ to minimize (2.2) are

$$
\begin{aligned}
&(-1)^m \lambda \hat{\beta}^{(m)}(x) + \frac{1}{n}\sum_{i=1}^{n} [b'(\eta(X_i;\hat{\beta})) - b'(\eta(X_i;\beta_0))] X_i^{(-m)}(t) \\
&\qquad = \frac{1}{n}\sum_{i=1}^{n} \sqrt{b''(\eta(X_i;\beta_0))}\ \epsilon_i X_i^{(-m)}(t) \qquad \text{(A.1)}
\end{aligned}
$$

and, for $k = 1, \ldots, m$,

$$
\frac{1}{n}\sum_{i=1}^{n} [b'(\eta(X_i;\hat{\beta})) - b'(\eta(X_i;\beta_0))] X_i^{(-k)}(1) = \frac{1}{n}\sum_{i=1}^{n} \sqrt{b''(\eta(X_i;\beta_0))}\ \epsilon_i X_i^{(-k)}(1). \qquad \text{(A.2)}
$$

In (A.1) and (A.2), we approximate $b'(\eta(X_i;\hat{\beta}))-b'(\eta(X_i;\beta_0))$ by $b''(\eta(X_i;\beta_0))\eta(X_i;\hat{\beta}-\beta_0)$, and consider a linear estimator $\tilde{\beta}$ that satisfies

$$\begin{aligned}&(-1)^m\lambda\tilde{\beta}^{(m)}(t)+\frac{1}{n}\sum_{i=1}^{n}b''(\eta(X_i;\beta_0))\eta(X_i;\tilde{\beta}-\beta_0)X_i^{(-m)}(t)\\&\quad=\frac{1}{n}\sum_{i=1}^{n}\sqrt{b''(\eta(X_i;\beta_0))}\,\epsilon_iX_i^{(-m)}(t),\end{aligned}\tag{A.3}$$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}b''(\eta(X_i;\beta_0))\eta(X_i;\tilde{\beta}-\beta_0)X_i^{(-k)}(1)\\&\quad=\frac{1}{n}\sum_{i=1}^{n}\sqrt{b''(\eta(X_i;\beta_0))}\,\epsilon_iX_i^{(-k)}(1),\quad k=1,\dots,m.\end{aligned}\tag{A.4}$$

Later, we show that $\hat{\beta}$ and $\tilde{\beta}$ are asymptotically equivalent in terms of the prediction error.

Let $Z(t)=\sqrt{b''(\eta(X;\beta_0))}\,X(t)$, and

$$\widetilde{Z}(1)=\left[Z^{(-1)}(1),Z^{(-2)}(1),\dots,Z^{(-m)}(1)\right]^T.$$

Then (A.3) and (A.4) are, respectively,

$$\begin{aligned}&(-1)^m\lambda\tilde{\beta}^{(m)}(t)+\frac{1}{n}\sum_{i=1}^{n}\eta(Z_i;\tilde{\beta}-\beta_0)Z_i^{(-m)}(t)\\&\quad=\frac{1}{n}\sum_{i=1}^{n}\epsilon_iZ_i^{(-m)}(t),\end{aligned}\tag{A.5}$$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\eta(Z_i;\tilde{\beta}-\beta_0)Z_i^{(-k)}(1)\\&\quad=\frac{1}{n}\sum_{i=1}^{n}\epsilon_iZ_i^{(-k)}(1),i=1,\dots,m.\end{aligned}\tag{A.6}$$

Let

$$\hat{H}=\frac{1}{n}\sum_{i=1}^{n}\widetilde{Z}_i(1)\widetilde{Z}_i(1)^T=\frac{1}{n}\sum_{i=1}^{n}b''(\eta(X_i;\beta_0))\widetilde{X}_i(1)\widetilde{X}_i(1)^T,$$

$$\hat{G}^{(-m\bullet)}(t)=\frac{1}{n}\sum_{i=1}^{n}Z_i^{(-m)}(t)\widetilde{Z}_i(1)=\frac{1}{n}\sum_{i=1}^{n}b''(\eta(X_i;\beta_0))X_i^{(-m)}(t)\widetilde{X}_i(1),$$

$$\hat{G}^{(-j,-k)}(t,s)=\frac{1}{n}\sum_{i=1}^{n}Z_i^{(-j)}(t)Z_i^{(-k)}(s)=\frac{1}{n}\sum_{i=1}^{n}b''(\eta(X_i;\beta_0))X_i^{(-j)}(t)X_i^{(-k)}(s),$$

for $j,k=0,1,\ldots,m$. Given $\hat{G}^{(-m\bullet)}$ and $\hat{H}$, take

$$\hat{U}(t;Z)=Z^{(-m)}(t)-\hat{G}^{(-m\bullet)}(t)^T\hat{H}^{-1}\widetilde{Z}(1).$$

Lemma A.2. The $\tilde{\beta}^{(m)}$ satisfy the differ-integral equation

$$\lambda\tilde{\beta}^{(m)}(t)+\int_0^1\hat{Q}(t,s)\Big(\tilde{\beta}^{(m)}(s)-\beta_0^{(m)}(s)\Big)ds=(-1)^m\frac{1}{n}\sum_{i=1}^n\epsilon_i\hat{U}(t;Z_i). \quad \text{(A.7)}$$

where

$$\hat{Q}(t,s)=\frac{1}{n}\sum_{i=1}^n Z_i^{(-m)}(t)Z_i^{(-m)}(s)-\hat{G}^{(-m\bullet)}(t)^T\hat{H}^{-1}\hat{G}^{(-m\bullet)}(s).$$

Lemma A.3. Let $\hat{Q}^+=(\lambda I+\hat{Q})^{-1}$. For any random function Z, we have

$$\begin{aligned}&\int_0^1 Z(t)\Big(\tilde{\beta}(t)-\beta_0(t)\Big)dt\\&=(-1)^{m+1}\lambda\int_0^1\hat{U}(t;Z)\hat{Q}^+\beta_0^{(m)}(t)dt+\frac{1}{n}\sum_{i=1}^n\epsilon_i\int_0^1\hat{U}(t;Z)\hat{Q}^+\hat{U}(t;Z_i)dt\\&\quad+\frac{1}{n}\sum_{i=1}^n\epsilon_i\widetilde{Z}(1)\hat{H}^{-1}\tilde{Z}_i(1). \quad \text{(A.8)}\end{aligned}$$

Recall that $Z_i(t)=\sqrt{b''(\eta(X_i;\beta_0))}\;X_i(t)$, and observe that $\hat{Q}(t,s)=1/n\sum_{i=1}^n\hat{U}(t;Z_i)\hat{U}(s,Z_i)$. Write the spectral expansion of $\hat{Q}$ as $\hat{Q}(t,s)=\sum_{j=1}^\infty\hat{\kappa}_j\hat{\phi}_j(t)\hat{\phi}_j(s)$, where $\hat{\kappa}_1\geq\hat{\kappa}_2\geq\cdots\geq 0$ are the eigenvalue sequence of the linear operator with kernel $\hat{Q}$, and $\hat{\phi}_1,\hat{\phi}_2,\cdots$ are the respective orthonormal eigenfunctions. Write

$$\hat{U}(t;Z_i)=\sum_{j=1}^\infty\hat{\xi}_j^{(i)}\hat{\phi}_j(t),\quad i=1,\ldots,n.$$

Since $\hat{Q}(t,s)=n^{-1}\sum_{i=1}^n\hat{U}(t;Z_i)\hat{U}(s;Z_i)$, we have

$$\hat{\kappa}_j=\frac{1}{n}\sum_{i=1}^n(\hat{\xi}_j^{(i)})^2,\qquad\frac{1}{n}\sum_{i=1}^n\hat{\xi}_j^{(i)}\hat{\xi}_k^{(i)}=0,\quad j\neq k.$$

Let $\Gamma(t,s)=E\left[b''(\eta(X;\beta_0))X(s)X(t)\right]$ and

$$\Gamma_n(t,s)=\frac{1}{n}\sum_{i=1}^n Z_i(t)Z_i(s)=\frac{1}{n}\sum_{i=1}^n b''(\eta(X_i;\beta_0))X_i(t)X_i(s).$$

It follows from (A.8) that

$$\|\tilde{\beta}-\beta_0\|_{\Gamma_n}^2=\frac{1}{n}\sum_{i=1}^n\Big\{\int_0^1 Z_i(t)(\tilde{\beta}(t)-\beta_0(t))dt\Big\}^2$$

$$
\begin{aligned}
&\le 3\lambda^2 \int\int \hat{Q}(t,s)\hat{Q}^+\beta_0^{(m)}(t)\hat{Q}^+\beta_0^{(m)}(s)dtds \\
&\quad + \frac{3}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\int_0^1\int_0^1 \hat{Q}(t,s)\hat{Q}^+\hat{U}(t;Z_i)\hat{Q}^+\hat{U}(s;Z_i)dtds \\
&\quad + \frac{3}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\tilde{Z}_i^T(1)\hat{H}^{-1}\tilde{Z}_j(1) \\
&= 3\lambda^2\sum_{j=1}^{\infty}\frac{\hat{\eta}_j^2\hat{\kappa}_j}{(\lambda+\hat{\kappa}_j)^2} + \frac{3}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\sum_{k=1}^{\infty}\frac{\hat{\xi}_k^{(i)}\hat{\xi}_k^{(j)}\hat{\kappa}_k}{(\lambda+\hat{\kappa}_k)^2} \\
&\quad + \frac{3}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\epsilon_i\epsilon_j\tilde{Z}_i^T(1)\hat{H}^{-1}\tilde{Z}_j(1), \qquad \text{(A.9)}
\end{aligned}
$$

where $\beta_0^{(m)}(t) = \sum_{j=1}^{\infty}\hat{\eta}_j\hat{\phi}_j(t)$ and $\sum_{j=1}^{\infty}\hat{\eta}_j^2 < \infty$. We find the asymptotic order of each term in (A.9). The first term is

$$
3\lambda^2\sum_{j=1}^{\infty}\frac{\hat{\eta}_j^2\hat{\kappa}_j}{(\lambda+\hat{\kappa}_j)^2} \le 3\lambda^2\sum_{j=1}^{\infty}\hat{\eta}_j^2\sup_j\frac{\hat{\kappa}_j}{(\lambda+\hat{\kappa}_j)^2} \le 3\lambda^2\sum_{j=1}^{\infty}\hat{\eta}_j^2\frac{1}{4\lambda} = O(\lambda).
$$

It is easy to see that the expectation of the third term in (A.9) is $(3\sigma^2/n)E[\tilde{Z}^T(1)\hat{H}^{-1}\tilde{Z}(1)]$, which is of order $O(n^{-1})$. Similarly, the expectation of the second term in (A.9) is

$$
\begin{aligned}
\frac{3\sigma^2}{n}E\Big\{\sum_{j=1}^{\infty}\frac{\hat{\kappa}_j^2}{(\lambda+\hat{\kappa}_j)^2}\Big\} &\le \frac{3\sigma^2}{n}E\Big\{\sum_{j=1}^{J}\frac{\hat{\kappa}_j^2}{(\lambda+\hat{\kappa}_j)^2} + \sum_{j=J+1}^{\infty}\frac{\hat{\kappa}_j^2}{(\lambda+\hat{\kappa}_j)^2}\Big\} \\
&\le \frac{3J\sigma^2}{n} + \frac{3\sigma^2}{n\lambda^2}E\Big\{\sum_{j=J+1}^{\infty}\hat{\kappa}_j^2\Big\}. \qquad \text{(A.10)}
\end{aligned}
$$

Next, we discuss the choice of J and the upper bound for (A.10). Let $\bar{U}(t;Z_i) = Z_i^{(-m)}(t) - C^{(-m\bullet)}(t)^T H^{-1}\widetilde{Z}_i(1)$, where $C^{(-m\bullet)}(t) = E[Z^{(-m)}(t)\widetilde{Z}(1)]$, $H = E[\widetilde{Z}(1)\widetilde{Z}(1)^T]$. Then, $E[\bar{U}(t;Z_i)\bar{U}(s;Z_i)]$ is

$$
Q(t,s) = G(t,s) - C^{(-m\bullet)}(t)^T H^{-1}C^{(-m\bullet)}(s), \qquad \text{(A.11)}
$$

where $G(t,s) = E(Z^{(-m)}(t)Z^{(-m)}(s))$.

Lemma A.4. The operator Q is nonnegative definite. If $\kappa_1 \ge \kappa_2 \ge \cdots > 0$ are the nonzero eigenvalues of Q and ϕ_j is the eigenfunction corresponding to κ_j, then, under A3, $\kappa_k \asymp k^{-2(m+r)}$.

Let

$$\hat{\Delta}_{jk} = \int_0^1 \int_0^1 (\hat{Q}(t,s) - Q(t,s))\phi_j(t)\phi_k(s)dtds.$$

Hall and Hosseini-Nasab (2006) have shown that

$$\hat{\kappa}_j - \kappa_j = \Delta_{jj} + \sum_{k:k\neq j} \frac{\Delta_{jk}^2}{(\kappa_j - \kappa_k)} + \text{remainder terms},$$

where the expected value of the remainder terms can be shown to be of order n^{-2} which is negligible in our analysis. Using the discussion in Section 5.3 of Hall and Horowitz (2007), we can show that $\mathbb{E}(\hat{\Delta}_{jj}^2) = O(n^{-1}\kappa_j^2)$ and

$$n\mathbb{E}\Big[\sum_{k:k\neq j} \frac{\Delta_{jk}^2}{(\kappa_j - \kappa_k)}\Big]^2 \leq C\kappa_j^2,$$

uniformly in j. Therefore,

$$\mathbb{E}\Big(\sum_{k=\varrho+1}^{\infty} \hat{\kappa}_k^2\Big) \leq (1 + O(n^{-1})) \sum_{k=\varrho+1}^{\infty} \kappa_k^2.$$

Choosing $J = \lambda^{-1/(2(m+r))}$, we have $E\Big\{\sum_{j=J+1}^{\infty} \hat{\kappa}_j^2\Big\} = O(\lambda^{2-1/[2(m+r)]})$. Combining these results,

$$E\|\tilde{\beta} - \beta_0\|_\Gamma^2 = O\Big(\lambda + n^{-1}\lambda^{-\frac{1}{2(m+r)}} + n^{-1}\Big).$$

Since b'' has a positive lower bound, we also have

$$\|\tilde{\beta} - \beta_0\|_K^2 = O_p\Big(\lambda + n^{-1}\lambda^{-\frac{1}{2(m+r)}} + n^{-1}\Big).$$

Lemma A.5. $\|\hat{\beta} - \beta_0\|_\Gamma^2 = \|\tilde{\beta} - \beta_0\|_\Gamma^2 + O_p\Big(\lambda + n^{-1}\lambda^{-1/[2(m+r)]} + n^{-1}\Big)$.

It follows from Lemma A.5 that $\|\hat{\beta} - \beta_0\|_\Gamma^2$ and $\|\tilde{\beta} - \beta_0\|_\Gamma^2$ have the same asymptotic order. This completes the proof of the theorem.

B. Proofs of Lemmas

Proof of Lemma A.1. Denote

$$L(\beta) = -\frac{1}{n}\sum_{i=1}^{n} \Big\{Y_i\eta(X_i;\beta) - b(\eta(X_i;\beta))\Big\} + \frac{\lambda}{2}\int_0^1 [\beta^{(m)}(t)]^2 dt.$$

For any $\delta > 0$ and $\beta_1 \in W_2^m$,

$$L(\beta + \delta\beta_1) - L(\beta_1) = -\frac{1}{n}\sum_{i=1}^{n} \Big\{\delta Y_i\eta(X_i;\beta_1) - [b(\eta(X_i;\beta + \delta\beta_1) - b(\eta(X_i;\beta))]\Big\}$$

$$+\delta\lambda \textstyle\int_0^1 \beta^{(m)}(t)\beta_1^{(m)}(t)dt + O(\delta^2)$$

So, by letting $\delta \to 0$, it is easy to see that the necessary and sufficient condition for β to minimize (2.2) is, for any $\beta_1 \in W_2^m$, $L_1(\beta, \beta_1) = 0$, where

$$L_1(\beta,\beta_1) = -\frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i\eta(X_i;\beta_1) - b'(\eta(X_i;\beta))\eta(X_i;\beta_1)\Big\} + \lambda\int_0^1 \beta^{(m)}(t)\beta_1^{(m)}(t)dt. \tag{B.1}$$

In (B.1), letting $\beta_1(t) = t^k$, $k = 0, 1, \ldots, m$, we obtain m equalities in (A.2). For example, when $\beta_1(t) = 1$, we have

$$\frac{1}{n}\sum_{i=1}^{n} b'(\eta(X_i;\beta))X_i^{(-1)}(1) = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{(-1)}(1).$$

Further, since

$$\eta(X_i; 1-t) = \int_0^1 X_i(t)\int_t^1 dsdt = \int_0^1 X_i^{(-1)}(t)dt = X_i^{(-2)}(1),$$

when $\beta_1(t) = t$,

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i\eta(X_i;t) - b'(\eta(X_i;\beta))\eta(X_i;t)\Big\}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\Big\{Y_i\eta(X_i;1-t) - b'(\eta(X_i;\beta))\eta(X_i;1-t)\Big\}\\
&\quad= \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{(-2)}(1) - \frac{1}{n}\sum_{i=1}^{n} b'(\eta(X_i;\beta))X_i^{(-2)}(1).
\end{aligned}$$

Hence,

$$\frac{1}{n}\sum_{i=1}^{n} b'(\eta(X_i;\beta))X_i^{(-2)}(1) = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{(-2)}(1).$$

Following the same procedure, it may be shown that (A.2) holds.

Next, using these equalities, we show that $L_1(\beta, \beta_1) = \int_0^1 L_2(\beta)\beta_1^{(m)}(t)dt$, where

$$L_2(\beta) = \lambda\beta^{(m)}(t) + (-1)^m\Big\{\frac{1}{n}\sum_{i=1}^{n} b'(\eta(X_i;\hat{\beta}))X_i^{(-m)}(t) - \frac{1}{n}\sum_{i=1}^{n} Y_i X_i^{(-m)}(t)\Big\}.$$

Note that

$$\int_0^1 X_i(s)\beta_1(s) = \beta_1(1)X_i^{(-1)}(1) - \int_0^1 X_i^{(-1)}(s)\beta'(s)ds$$

$$
\begin{aligned}
&= \beta_1(1)X_i^{(-1)}(1) - \beta'(1)X_i^{(-2)}(1) + \int_0^1 X_i^{(-2)}(s)\beta''(s)ds \\
&= \quad \vdots \\
&= \sum_{k=0}^{m-1}(-1)^k\beta_1^{(k)}(1)X_i^{(-k-1)} + (-1)^m\int_0^1 X_i^{(-m)}(s)\beta^{(m)}(s)ds.
\end{aligned}
$$

Plugging this into $L_1(\beta,\beta_1) = 0$, together with (A.2), we have $L_1(\beta,\beta_1) = \int_0^1 L_2(\beta)\beta_1^{(m)}(t)dt$. Finally, since $L_1(\beta,\beta_1) = 0$ for any $\beta_1 \in W_2^m$, we have $L_2(\beta) = 0$ a.e., which completes the proof of the lemma.

Proof of Lemma A.2. Observe that

$$
\int_0^1 Z_i(s)\tilde{\beta}(s)ds = \sum_{k=0}^{m-1}(-1)^k\hat{\beta}^{(k)}(1)Z_i^{(-k-1)}(1) + (-1)^m\int_0^1 Z_i^{(-m)}(s)\hat{\beta}^{(m)}(s)ds.
$$

Hence, for $j = 1,\ldots,m$,

$$
\begin{aligned}
\int_0^1 \hat{G}^{(-j,0)}(1,s)\tilde{\beta}(s)ds = &\sum_{k=0}^{m-1}(-1)^k\tilde{\beta}^{(k)}(1)\hat{G}^{(-j,-k)}(1,1) \\
&+(-1)^m\int_0^1 \hat{G}^{(-j,-m)}(1,s)\tilde{\beta}^{(m)}(s)ds.
\end{aligned}
$$

From (A.6), we have

$$
\hat{H}\tilde{\beta}_v(1) = \frac{1}{n}\sum_{i=1}^n \epsilon_i\widetilde{Z}(1) - (-1)^m\int_0^1 \hat{G}^{(-m\bullet)}(s)\tilde{\beta}^{(m)}(s)ds, \tag{B.2}
$$

where $\tilde{\beta}_v(1) = \Big[\tilde{\beta}(1), -\tilde{\beta}'(1), \cdots, (-1)^{m-1}\tilde{\beta}^{(m-1)}(1)\Big]^T$. Hence, (A.7) follows by plugging (B.2) into (A.5).

Proof of Lemma A.3. Direct calculation yields

$$
\frac{1}{n}\sum_{i=1}^n \epsilon_i\tilde{Z}(1) = \hat{H}\beta_{0v}(1) + (-1)^m\int_0^1 \hat{G}^{(-m\bullet)}(s)\beta_0^{(m)}(s)ds + \frac{1}{n}\sum_{i=1}^n \epsilon_i\widetilde{Z}_i(1).
$$

Combining this with (B.2) leads

$$
\begin{aligned}
&\tilde{\beta}_v(1) - \beta_{0v}(1) \\
&= (-1)^{m+1}\int_0^1 \hat{H}^{-1}\hat{G}^{(-m\bullet)}(s)\Big(\tilde{\beta}^{(m)}(s) - \beta_0^{(m)}(s)\Big)ds + \frac{1}{n}\sum_{i=1}^n \epsilon_i\hat{H}^{-1}\widetilde{Z}_i(1).
\end{aligned}
$$

Therefore,

$$\begin{aligned}
&\int_0^1 Z(s)\Big(\hat{\beta}(s)-\beta_0(s)\Big)ds \\
&= \widetilde{Z}(1)^T\Big(\tilde{\beta}_v(1)-\beta_{0v}(1)\Big)+(-1)^m\int_0^1 Z^{(-m)}(s)\Big(\tilde{\beta}^{(m)}(s)-\beta_0^{(m)}(s)\Big)ds \\
&= (-1)^m\int_0^1 \hat{U}(s;Z)\ \Big(\tilde{\beta}^{(m)}(s)-\beta_0^{(m)}(s)\Big)ds+\frac{1}{n}\sum_{i=1}^n \epsilon_i\widetilde{Z}(1)^T\hat{H}^{-1}\widetilde{Z}_i(1). \quad \text{(B.3)}
\end{aligned}$$

From Lemma A.2.,

$$\tilde{\beta}^{(m)}-\beta_0^{(m)} = -\lambda\hat{Q}^+\beta_0^{(m)}+(-1)^m\frac{1}{n}\sum_{i=1}^n\epsilon_i\hat{Q}^+\hat{U}(t;Z_i).$$

Plugging this into (B.3) leads to (A.8). This completes the proof of the lemma.

Proof of Lemma A.4 Let $H(K)$ denote the reproducing kernel space associated withe the kernel K. For two covariance kernel K and L on $[0,1]^2$ we write $K \ll L$ if $cL-K$ is nonnegative definite for some positive constant c. Then, $K \ll L$ implies $H(K)\subset H(L)$. From (A.11),

$$Q \ll G \Longleftrightarrow H(Q)\subset H(G).$$

Let $\lambda_1(G)\geq\lambda_2(G)\geq\cdots>0$ be the eigenvalues of G. Let ϕ_k be the eigenfunction of Q which corresponds to κ_k such that $Q\phi_k=\kappa_k\phi_k$. The minimax principle [see Weidmann (1980, Thm. 7.3) yields

$$Q \ll G \Longrightarrow \kappa_k \leq c\lambda_k(G) \text{ for some positive constant } c. \quad \text{(B.4)}$$

We may write $H(G)=H(Q)\oplus H(G-Q)$, and $H(G-Q)$ is the orthogonal complement of $H(Q)$. Note that $H(G-Q)$ is a finite dimensional space with rank m. Let $f_1,\ldots,f_m$ be an orthonormal base of $H(G-Q)$, and let $\mathcal{F}^\perp$ denote the sets of normalized functions in L_2 that are orthogonal to $f_1,\ldots,f_m$. Let $\Xi_k^\perp=\text{span}\{\phi_k,\phi_{k+1},\ldots,\}$. The minimax principle implies

$$\lambda_{k+m}(G)\leq \sup_{f\in\mathcal{F}^\perp\cap\Xi_k^\perp}\langle Gf,f\rangle = \sup_{f\in\mathcal{F}^\perp\cap\Xi_k^\perp}\langle Qf,f\rangle \leq \sup_{f\in\Xi_k^\perp}\langle Qf,f\rangle=\kappa_k. \quad \text{(B.5)}$$

From A1, Ritter et al. (1995) showed that $\lambda_k(G)\asymp k^{-2(m+r)}$. Further, (B.4) and (B.5) yield that $\kappa_k\asymp k^{-2(m+r)}$. Since $H(G-Q)$ is not an empty set, the operator Q is not a strictly positive definite operator, and Q has m zero eigenvalues.

Proof of Lemma A.5. In the lemma, we discuss the relationship between $\hat{\beta}$ and $\tilde{\beta}$. Denote

$$\Delta(\beta_1,\beta_2;X_i)=b'(\eta(X_i;\beta_1))-b'(\eta(X_i;\beta_2))-b''(\eta(X_i;\beta_2))\eta(X_i;\beta_1-\beta_2).$$

Let δ_1 satisfy

$$(-1)^m\lambda\delta_1^{(m)}(t)+\frac{1}{n}\sum_{i=1}^n b''(\eta(X_i;\tilde\beta))\eta(X_i;\delta_1)X_i^{(-m)}(t)$$
$$=-\frac{1}{n}\sum_{i=1}^n X_i^{(-m)}(t)\Delta(\tilde\beta,\beta_0;X_i).$$

For $k\ge 2$, let δ_k satisfy

$$(-1)^m\lambda\delta_k^{(m)}(t)+\frac{1}{n}\sum_{i=1}^n b''(\eta(X_i;\delta_{k-1}))\eta(X_i;\delta_k)X_i^{(-m)}(t)$$
$$=-\frac{1}{n}\sum_{i=1}^n X_i^{(-m)}(t)\Delta(\tilde\beta+\sum_{j=1}^{k-1}\delta_j,\tilde\beta+\sum_{j=1}^{k-2}\delta_j;X_i).$$

By summing all these equations together, it is easy to verify that $\hat\beta=\tilde\beta+\sum_{k=1}^\infty\delta_k$. Following the same discussion in Lemma A.1, for any $\beta_1\in W_2^m$,

$$\frac{1}{n}\sum_{i=1}^n\Big\{b''(\eta(X_i;\tilde\beta))\eta(X_i;\delta_1)\eta(X_i;\beta_1)-\sqrt{b''(\eta(X_i;\tilde\beta))}\eta(X_i;\beta_1)\Delta_i\Big\}$$
$$+\lambda\int_0^1\delta_1^{(m)}(t)\beta_1^{(m)}(t)dt=0,$$

where $\Delta_i=\Delta(\tilde\beta,\beta_0;X_i)$. By choosing $\beta_1=\delta_1$,

$$\frac{1}{n}\sum_{i=1}^n b''(\eta(X_i;\tilde\beta))[\eta(X_i;\delta_1)]^2-\frac{1}{n}\sum_{i=1}^n\sqrt{b''(\eta(X_i;\tilde\beta))}\eta(X_i;\beta_1)\Delta_i$$
$$+\lambda\int_0^1[\delta_1^{(m)}(t)]^2dt=0.$$

Therefore,

$$\|\delta_1\|_{\Gamma_n}^2\le\frac{c_2}{nc_1}\sum_{i=1}^n b''(\eta(X_i;\tilde\beta))[\eta(X_i;\delta_1)]^2+\lambda\int_0^1[\delta_1^{(m)}(t)]^2dt$$
$$\le\frac{c_2}{nc_1}\sum_{i=1}^n\sqrt{b''(\eta(X_i;\tilde\beta))}\eta(X_i;\delta_1)\Delta_i$$
$$\le C_1\|\delta_1\|_{\Gamma_n}\Big\{\frac{1}{n}\sum_{i=1}^n[\eta(X_i;\tilde\beta-\beta)]^4\Big\}^{1/2}.$$

Recall that the (ν_j,ψ_j) are the eigenvalue-eigenfunction pairs for the covariance kernel K. By A4,

$$E\Big\{\frac{1}{n}\sum_{i=1}^n[\eta(X_i;\tilde\beta-\beta)]^4\Big|\tilde\beta\Big\}$$

$$
\begin{aligned}
&= E\Big\{\sum_{j_1}^{\infty}\sum_{j_2}^{\infty}\sum_{j_3}^{\infty}\sum_{j_4}^{\infty}\zeta_{j_1}\zeta_{j_2}\zeta_{j_3}\zeta_{j_4}\eta(\psi_{j_1},\tilde{\beta}-\beta)\eta(\psi_{j_2},\tilde{\beta}-\beta)\eta(\psi_{j_3},\tilde{\beta}-\beta)\eta(\psi_{j_4},\tilde{\beta}-\beta)\Big|\tilde{\beta}\Big\}\\
&= \sum_{j=1}^{\infty} E(\zeta_j^4)[\eta(\psi_j,\tilde{\beta}-\beta)]^4 + \sum_{j_1}\sum_{j_2\neq j_1}\nu_{j_1}\nu_{j_2}[\eta(\psi_{j_1},\tilde{\beta}-\beta)]^2[\eta(\psi_{j_2},\tilde{\beta}-\beta)]^2\\
&\le C_2\sum_{j=1}^{\infty}\nu_j^2[\eta(\psi_j,\tilde{\beta}-\beta)]^4 + \sum_{j_1}\sum_{j_2\neq j_1}\nu_{j_1}\nu_{j_2}[\eta(\psi_{j_1},\tilde{\beta}-\beta)]^2[\eta(\psi_{j_2},\tilde{\beta}-\beta)]^2\\
&\le (1+C_2)\sum_{j_1=1}^{\infty}\nu_{j_1}[\eta(\psi_{j_1},\tilde{\beta}-\beta)]^2\sum_{j_2=1}^{\infty}\nu_{j_2}[\eta(\psi_{j_2},\tilde{\beta}-\beta)]^2\\
&= (1+C_2)\|\tilde{\beta}-\beta\|_K^4.
\end{aligned}
$$

So,

$$E\|\delta_1\|_\Gamma^2 = E\|\delta_1\|_{\Gamma_n}^2 \le C_2\Big(\lambda + n^{-1}\lambda^{-\frac{1}{2(m+r)}} + n^{-1}\Big)^2.$$

Similarly, we may establish, for $k \ge 1$,

$$E\|\delta_k\|_\Gamma^2 = E\|\delta_1\|_{\Gamma_n}^2 \le C_2\Big(\lambda + n^{-1}\lambda^{-\frac{1}{2(m+r)}} + n^{-1}\Big)^{2k}.$$

Therefore, $\sum_{k=1}^{\infty}\delta_k$ is of order $O_p(\lambda + n^{-1}\lambda^{-1/[2(m+r)]} + n^{-1})$. This shows the lemma.

References

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159-2179.

Cai, T. T. and Yuan, M. (2012). Minimax and adaptive prediction for functional linear regression. *J. Amer. Statist. Assoc.* **107**, 1201-1216.

Cardot, H. and Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92**, 24-41.

Carroll, R. J. and Ruppert, D. (2006). Discussion on "Conditional growth charts" by wei and he. *Ann. Statist.* **34**, 2098-2104.

Crainiceanu, C. M., Staicu, A.-M. and Di, C.-Z. (2009). Generalized multilevel functional regression. *J. Amer. Statist. Assoc.* **104**, 1550-1561.

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.

Dou, W., Pollard, D. and Zhou, H. H. (2010). Functional regression for general exponential families. Technical report, Department of Statistics, Yale University, New Haven, CT.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer-Verlag Inc, New York.

Goldsmith, J., Bobb, J., Crainiceanu, C. M., Caffo, B. and Reich, D. (2010). Penalized functional regression. *J. Comput. Graph. Statist.* **20**, 830-851.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

Hall, P. and Horowitz, J. L. (2007). Methodology and covergence rates for functional linear regression. *Ann. Statist.* **35**, 70-91.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis. *J. Roy. Statist. Soc. Ser. B* **68**, 109-126.

James, G. M. (2002). Generalized linear models with functional predictors. *J. Roy. Statist. Soc. Ser. B* **64**, 411-432.

James, G. M. and Hastie, T. J. (2001). Functional linear discriminant analysis for irregularly sampled curves. *J. Roy. Statist. Soc. Ser. B* **63**, 533-550.

Li, Y., Wang, N., and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions. *J. Amer. Statist. Assoc.* **105**, 621-633.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **33**, 774-805.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer-Verlag Inc, New York.

Ritter, K., Wozniakowski, H., Wasilkowski, G. W., and Woźniakowski, H. (1995). Multivariate integration and approximation for random fields satisfying Sacks-Ylvisaker conditions. *Ann. Appl. Probab.* **5**, 518-540.

Sacks, J. and Ylvisaker, D. (1966). Designs for regression problems with correlated errors. *Annnals of Mathematical Statistics* **37**, 66-89.

Sacks, J. and Ylvisaker, D. (1968). Designs for regression problems with correlated errors; many parameters. *Ann. Math. Statist.* **39**, 49-69.

Sacks, J. and Ylvisaker, D. (1970). Designs for regression problems with correlated errors, III. *Ann. Math. Statist.* **41**, 2057-2074.

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

Wahba, G. (1990). *Spline Models for Observational Data*, Volume 59 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Weidmann, J. (1980). *Linear operators in Hilbert spaces*. Springer-Verlag, New York.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica* **6**, 675-692.

Yuan, M. and Cai, T. T. (2010). A reproducing kernel Hilbert space approach to functional linear regression. *Ann. Statist.* **38**, 3412-3444.

Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA.

E-mail: pangdu@vt.edu

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

E-mail: wangxiao@purdue.edu

(Received August 2012; accepted June 2013)