Statistica Sinica **28** (2018), 2455-2468
doi:https://doi.org/10.5705/ss.202017.0099

ADAPTIVE FUNCTIONAL LINEAR REGRESSION VIA FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS AND BLOCK THRESHOLDING*

T. Tony Cai[1], Linjun Zhang[1] and Harrison H. Zhou[2]

[1] *University of Pennsylvania and* [2] *Yale University*

Abstract: Theoretical results in the functional linear regression literature have so far focused on minimax estimation where smoothness parameters are assumed to be known and the estimators typically depend on these smoothness parameters. In this paper we consider adaptive estimation in functional linear regression. The goal is to construct a single data-driven procedure that achieves optimality results simultaneously over a collection of parameter spaces. Such an adaptive procedure automatically adjusts to the smoothness properties of the underlying slope and covariance functions. The main technical tools for the construction of the adaptive procedure are functional principal component analysis and block thresholding. The estimator of the slope function is shown to adaptively attain the optimal rate of convergence over a large collection of function spaces.

Key words and phrases: Adaptive estimation, block thresholding, eigenfunction, eigenvalue, functional data analysis, functional principal component analysis, minimax estimation, rate of convergence, slope function, smoothing, spectral decomposition.

1. Introduction

Due to advances in technology, functional data now commonly arises in many different fields of applied sciences including, for example, chemometrics, biomedical studies, and econometrics. There has been extensive recent research on functional data analysis. Much progress has been made on developing methodologies for analyzing functional data. The two monographs by Ramsay and Silverman (2002, 2005) provide comprehensive discussions on the methods and applications. See also Ferraty and Vieu (2006).

Among many problems involving functional data, functional linear regression has received substantial attention. Consider a functional linear model where one

*In Memory of Peter G. Hall.

observes a random sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$ with

$$Y_i = a + \int_0^1 X_i(t)b(t)dt + Z_i, \tag{1.1}$$

where the response Y_i and the intercept a are scalars, the predictor X_i and slope function b are functions in $L_2([0,1])$, and the errors Z_i are independent and identically distributed $N(0, \sigma^2)$ variables. The goal is to estimate the slope function $b(t)$ and the intercept a based on the sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$. Note that once an estimator $\hat{b}$ of b is constructed, the intercept a can be estimated easily by

$$\hat{a} = \bar{Y} - \int_0^1 \bar{X}(t)\hat{b}(t)dt,$$

where $\bar{Y}$ and $\bar{X}$ are the averages of Y_i and X_i respectively. We shall thus focus our discussion in this paper on estimating the slope function b. The slope function is of significant interest on its own right. For example, knowing where b takes large or small values provides information about where a future observation x of X will have greatest leverage on the conditional mean of y given $X = x$.

The problem of slope-function estimation is intrinsically nonparametric and the convergence rate under the mean integrated squared error (MISE)

$$R(\hat{b}, b) = \mathbb{E}\|\hat{b} - b\|_2^2 = \mathbb{E}\int_0^1 \left\{\hat{b}(t) - b(t)\right\}^2 dt \tag{1.2}$$

is typically slower than n^{-1}. Rates of convergence of an estimator $\hat{b}$ to b have been studied in, e.g., Ferraty and Vieu (2000); Cuevas, Febrero and Fraiman (2002); Cardot and Sarda (2006); Li and Hsing (2007); Hall and Horowitz (2007). In particular, Hall and Horowitz (2007) showed that the minimax rate of convergence for estimating b under the MISE (1.2) is determined by the smoothness of the slope function, and of the covariance function for the distribution of explanatory variables. Cai and Hall (2006) considered a related prediction problem and Müller and Stadtmüller (2005) studied generalized functional linear models.

The theory on slope function estimation has so far focused on the minimax estimation where these smoothness parameters are assumed to be known. The estimators typically depend on the smoothness parameters. Although minimax risk provides a useful uniform benchmark for the comparison of estimators, minimax estimators often require full knowledge of the parameter space which is unknown in practice. A minimax estimator designed for a specific parameter space typically performs poorly over another parameter space. This makes adaptation essential for functional linear regression.

In the present paper we consider adaptive estimation of the slope function b. The goal is to construct a single data-driven procedure that achieves optimality results simultaneously over a collection of parameter spaces. Such an adaptive procedure does not require the knowledge of the parameter space and automatically adjusts to the smoothness properties of the underlying slope and covariance functions. In Section 2, we construct a procedure for estimating the slope function b using functional principal component analysis (PCA) and block thresholding. The estimator is shown to adaptively achieve the optimal rate of convergence simultaneously over a collection of function classes.

The main technical tools are functional principal component analysis (PCA) and block thresholding. Functional PCA is a convenient and commonly used technique in functional data analysis. See, e.g., Ramsay and Silverman (2002, 2005). Block thresholding was first developed in nonparametric function estimation. It increases estimation precision and achieves adaptivity by utilizing information about neighboring coordinates. The idea of block thresholding can be traced back to Efromovich (1985) in estimating a density function using the trigonometric basis. It is further developed in wavelet function estimation. See Hall, Kerkyacharian and Picard (1998) for density estimation and Cai (1999) for nonparametric regression. Cai, Low and Zhao (2009) used weakly geometrically growing block size for sharp adaptation over ellipsoids in the context of the white noise model. In this paper we shall follow the ideas in Cai, Low and Zhao (2009) and use weakly geometrically growing block size for adaptive functional linear regression. Our results show that block thresholding naturally connects shrinkage rules developed in the classical normal decision theory with functional linear regression.

The proposed block thresholding procedure is easily implementable. A simulation study is carried out to investigate its numerical performance. In particular, we compare its finite-sample properties with those of the non-adaptive procedure introduced in Hall and Horowitz (2007). The results demonstrate the advantage of the proposed procedure.

The paper is organized as follows. In Section 2, after basic notations and facts on the spectral decomposition of the covariance function are reviewed, the block thresholding procedure for estimating the slope function b is defined in Section 2.2. Section 3 investigates the theoretical properties of the block thresholding procedure. It is shown that the estimator enjoys a high degree of adaptivity. Section 4 discusses the numerical performance of the proposed estimator and shows the advantage of the adaptive procedure. All the technical proofs are

given in the supplement (Cai, Zhang and Zhou (2017)).

2. Methodology

Estimating the slope function b in function linear regression involves solving an ill-posed inverse problem. The main difference with the conventional linear inverse problems is that the operator is not given in the functional linear regression. A major technical step in the construction of the slope function estimator is to estimate the eigenvalues and eigenfunctions of the unknown linear operator and to bound the errors between the estimates and the estimands. Necessary technical tools for slope function estimation include functional analysis and statistical smoothing. Specifically, our estimator is based on the functional principal component analysis and block thresholding techniques. In this section we will begin with spectral decomposition of the covariance function in terms of eigenvalues and eigenfunctions. We then introduce in Section 2.2 a blockwise James-Stein procedure to estimate the slope function b.

2.1. Spectral decomposition

Suppose we observe a random sample $\{(X_i, Y_i) : i = 1, \ldots, n\}$ as in (1.1). Let (X, Y, Z) denote a generic (X_i, Y_i, Z_i). Define the covariance function and the empirical covariance function respectively as

$$K(u,v) = \operatorname{cov}(X(u), X(v)),$$
$$\hat{K}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{X_i(u) - \bar{X}(u)\}\{X_i(v) - \bar{X}(v)\},$$

where $\bar{X} = (1/n)\sum X_i$. The covariance function K defines a linear operator which maps a function f to Kf given by $(Kf)(u) = \int K(u,v)f(v)dv$. We shall assume that the linear operator with kernel K is positive definite.

Write the spectral decompositions of the covariance functions K and $\hat{K}$ as

$$K(u,v) = \sum_{j=1}^{\infty}\theta_j\phi_j(u)\phi_j(v), \quad \hat{K}(u,v) = \sum_{j=1}^{\infty}\hat{\theta}_j\hat{\phi}_j(u)\hat{\phi}_j(v), \tag{2.1}$$

where

$$\theta_1 > \theta_2 > \cdots > 0, \text{ and } \hat{\theta}_1 \geq \hat{\theta}_2 \geq \cdots \geq \hat{\theta}_{n+1} = \cdots = 0 \tag{2.2}$$

are respectively the ordered eigenvalue sequences of the linear operators with kernels K and $\hat{K}$, and $\{\phi_j\}$ and $\{\hat{\phi}_j\}$ are the corresponding orthonormal eigenfunction sequences. The sequences $\{\phi_j\}$ and $\{\hat{\phi}_j\}$ each forms an orthonormal basis in $L_2([0,1])$.

The functional linear model (1.1) can be rewritten as

$$Y_i = \mu + \int \{X_i - \mathbb{E}(X)\} b + Z_i,\ i = 1, 2, \ldots, n \tag{2.3}$$

where $\mu = \mathbb{E}(Y_i) = a + \mathbb{E}\int Xb$. The Karhunen-Loève expansion of the random function $X_i - \mathbb{E}X$ is given by

$$X_i - \mathbb{E}X = \sum_{j=1}^{\infty} x_{i,j}\phi_j, \tag{2.4}$$

where the random variable $x_{i,j} = \int (X_i - \mathbb{E}X)\phi_j$ has mean zero and variance $\mathbb{V}ar(x_{i,j}) = \theta_j$. In addition, the random variables $x_{i,j}$ are uncorrelated. Expand the slope function b in the orthonormal basis $\{\phi_j\}$ as $b = \sum_{j=1}^{\infty} b_j\phi_j$. Then the model (2.3) can be written as

$$Y_i = \mu + \sum_{j=1}^{\infty} x_{i,j}b_j + Z_i,\ i = 1, 2, \ldots, n \tag{2.5}$$

and the problem of estimating the slope function b is transformed into the one of estimating the coefficients $\{b_j\}$ as well as the eigenfunctions $\{\phi_j\}$. Note that in (2.5) μ and $x_{i,j}$ are unknown, and thus need to be estimated from the data.

The mean μ of Y can be estimated easily by the sample mean $\hat{\mu} = \bar{Y}$. To estimate the $x_{i,j}$, we expand $X_i - \bar{X}$ in the orthonormal basis $\{\hat{\phi}_j\}$ as

$$X_i - \bar{X} = \sum_{j=1}^{n} \hat{x}_{i,j}\hat{\phi}_j \quad \text{for } i = 1, 2, \ldots, n, \tag{2.6}$$

where the random variables $\hat{x}_{i,j} = \int (X_i - \bar{X})\hat{\phi}_j$. Note that

$$\sum_{i=1}^{n} \hat{x}_{i,j} = \sum_{i=1}^{n} \int (X_i - \bar{X})\hat{\phi}_j = \int \left\{ \sum_{i=1}^{n} (X_i - \bar{X}) \right\} \hat{\phi}_j = 0$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i,j}\hat{x}_{i,k} = \iint \hat{K}(u, v)\hat{\phi}_j(u)\hat{\phi}_k(v) = \hat{\theta}_j\delta_{j,k} \tag{2.7}$$

for all j and k, where $\delta_{j,k}$ is the Kronecker delta with $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise. Since $\bar{Y} = a + \int_0^1 \bar{X}(t)b(t)dt + \bar{Z}$, we have

$$Y_i - \bar{Y} = \int \left(X_i - \bar{X}\right) b + Z_i - \bar{Z},\ i = 1, 2, \ldots, n.$$

Hence

$$Y_i - \bar{Y} = \sum_{j=1}^{n} \hat{x}_{i,j}\check{b}_j + Z_i - \bar{Z},\ i = 1, 2, \ldots, n, \tag{2.8}$$

where $\check{b}_j = \int b\hat{\phi}_j$, and consequently $b = \sum_{j=1}^{\infty} \check{b}_j \hat{\phi}_j$. Since the slope function b is unknown, the coefficients $\check{b}_j$ are also unknown and need to be estimated. A typical principal components regression approach is to replace "n" in (2.8) by a constant $m < n$ and estimate $\check{b}_j$ by ordinary least squares.

Since the "predictors" $\{\hat{x}_{i,j}\}_{1\le j\le n}$ in (2.8) are orthogonal to each other and $\sum_{i=1}^{n} \hat{x}_{i,j}^2 = \hat{\theta}_j n$ from (2.7), for $\hat{\theta}_j \neq 0$ we may estimate $\check{b}_j$ (or b_j) by

$$\begin{aligned}\tilde{b}_j &= \hat{\theta}_j^{-1} n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \hat{x}_{i,j} \\ &= \hat{\theta}_j^{-1} n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \int \left\{ X_i(u) - \bar{X}(u) \right\} \hat{\phi}_j(u) \\ &= \hat{\theta}_j^{-1} \int \hat{g}(u) \hat{\phi}_j(u) = \hat{\theta}_j^{-1} \hat{g}_j, \end{aligned} \tag{2.9}$$

where

$$\hat{g}(u) = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \left\{ X_i(u) - \bar{X}(u) \right\} \text{ and } \hat{g}_j = \int \hat{g}\hat{\phi}_j. \tag{2.10}$$

It is expected that $\hat{g}$ is approximately

$$g(u) = \mathbb{E}\left((Y - \mu) \left[X(u) - \mathbb{E}\{X(u)\} \right] \right) = \int K(u, v) b(v) \, dv.$$

Write $g = \sum_{j=1}^{\infty} g_j \phi_j$. It is easy to check that $b_j = \theta_j^{-1} g_j$. So the estimator $\tilde{b}_j = \hat{\theta}_j^{-1} \hat{g}_j$ in (2.9) can be regarded as an empirical version of the true coefficient b_j. We shall construct an adaptive estimator of b_j based on the empirical coefficients $\tilde{b}_j$ by using a block thresholding technique.

2.2. A block thresholding procedure

Block thresholding techniques have been well developed in nonparametric function estimation literature. See, e.g., Efromovich (1985), Hall, Kerkyacharian and Picard (1998) and Cai (1999). In this paper we shall use a block thresholding method with weakly geometrically growing block size for adaptive functional linear regression. This method was used in Cai, Low and Zhao (2009) for sharp adaptive estimation over ellipsoids in the classical white noise model.

The block thresholding procedures work especially well with homoscedastic Gaussian data. However, in the current setting the empirical coefficients $\tilde{b}_j$ are heteroscedastic with growing variances. We will see in Lemma 3 in Section S1 that the variance of $\tilde{b}_j$ is approximately $\sigma^2 \theta_j^{-1} n^{-1}$, getting large as j increases. We shall thus rescale the $\tilde{b}_j$ to stabilize the variances.

With this notation the block thresholding procedure can then be described in detail as follows. Let

$$\hat{m}^* = \arg\min\left(m : \frac{\hat{\theta}_m}{\hat{\theta}_1} \le n^{-1/3}\right). \tag{2.11}$$

It will be shown in Section S1 that there is no need ever to go beyond the $\hat{m}^*$-th term under certain regularity conditions. We define

$$\tilde{g}_j = \begin{cases} \hat{g}_j & j < \hat{m}^*, \\ 0 & \text{otherwise}, \end{cases}$$

and set

$$\tilde{d}_j = \hat{\theta}_j^{-1/2}\tilde{g}_j \quad \text{and} \quad d_j = \theta_j^{-1/2} g_j. \tag{2.12}$$

Lemma 4 in Section S1 shows that the variance $\mathbb{V}ar(\tilde{d}_j) = \sigma^2/n\{1 + o(1)\}$ and so the $\tilde{d}_j$ are nearly homoscedastic. We shall apply a blockwise James-Stein procedure to $\tilde{d}_j$ to construct an estimator $\hat{d}_j$ of d_j and then estimate the b_j by $\hat{b}_j = \hat{\theta}_j^{-1/2}\hat{d}_j$.

The block thresholding procedure for estimating the slope function b has three steps.

1. Divide the indices $\{1, 2, \ldots, \hat{m}^*\}$ into nonoverlapping blocks $B_1, B_2, \ldots, B_N$ with $\text{Card}(B_i) = \left\lfloor (1 + 1/\log n)^{i+1} \right\rfloor$.

2. Apply a blockwise James-Stein rule to each block. For all $j \in B_i$, set

$$\hat{d}_j = \left(1 - \frac{2L_i\sigma^2}{nS_i^2}\right)_+ \cdot \tilde{d}_j, \tag{2.13}$$

   where $S_i^2 = \sum_{j \in B_i} \tilde{d}_j^2$ and $L_i = \text{Card}(B_i)$.

3. Set $\hat{b}_j = \hat{\theta}_j^{-1/2}\hat{d}_j$. The estimator of b is then given by

$$\hat{b}(u) = \sum_{j=1}^{\hat{m}^*} \hat{b}_j\hat{\phi}_j(u) = \sum_{j=1}^{\hat{m}^*} \rho_j\tilde{b}_j\hat{\phi}_j(u), \tag{2.14}$$

   where $\rho_j = (1 - 2L_j\sigma^2/nS_i^2)_+$ for all $j \in B_i$ is the shrinkage factor.

The block thresholding procedure given above is purely data-driven and is easy to implement. In particular it does not require the knowledge of the rate of decay of the eigenvalues θ_j or the coefficients b_j of the slope function b. In contrast, the minimax rate optimal estimator given in Hall and Horowitz (2007) critically depends on the rates of decay of θ_j and b_j.

Remark 1. We have used the blockwise James-Stein procedure in (2.13) because of its simplicity. In addition to the James-Stein rule, other shrinkage rules such as the blockwise hard thresholding rule

$$\hat{d}_j = \tilde{d}_j \cdot I\left(S_i^2 \geq \frac{\lambda L_i \sigma^2}{n}\right)$$

can be used as well.

Remark 2. In the procedure we assume σ is known, since it can be estimated easily. In (2.8), we may apply principal components regression by replacing "n" in (2.8) with a constant $m = \log^2 n$. Let $\hat{b}_j$ be the ordinary least squares estimate of $\check{b}_j$. It can be shown easily that $\sum_{j=1}^{m} \hat{\mathrm{b}}_j \hat{\phi}_j$ is a consistent estimate of b. Then we obtain a consistent estimate of σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y} - \sum_{j=1}^{m} \hat{x}_{i,j}\hat{\mathrm{b}}_j\right)^2.$$

3. Theoretical Properties

We now turn to the asymptotic properties of the block thresholding procedure for the functional linear regression under the mean integrated squared error (1.2). The theoretical results show that the block thresholding estimator given in (2.14) adaptively attains the exact minimax rate of convergence simultaneously over a large collection of function spaces.

In this section we shall begin by considering adaptivity of the block thresholding estimator over the following function spaces which have been considered by Cai and Hall (2006) and Hall and Horowitz (2007) in the contexts of prediction and slope function estimation. These function classes arise naturally in functional linear regression based on functional principal component analysis. For more details, see Cai and Hall (2006) and Hall and Horowitz (2007). See also Hall and Hosseini-Nasab (2006).

Let $\beta > 0$ and $M_* > 0$ be constants. Define the function class for b by

$$\mathcal{B}^{\beta}(M_*) = \left\{ b = \sum_{j=1}^{\infty} b_j \, \phi_j, \text{ with } |b_j| \leq M_* \, j^{-\beta} \text{ for } j = 1, 2, \ldots \right\}. \tag{3.1}$$

We can interpret this as a "smoothness class" of functions, where the functions become "smoother" (measured in the sense of generalized Fourier expansions in the basis $\{\phi_j\}$) as β increases. We shall also assume the eigenvalues satisfy

$$M_0^{-1} \, j^{-\alpha} \leq \theta_j \leq M_0 \, j^{-\alpha}, \quad \theta_j - \theta_{j+1} \geq M_0^{-1} \, j^{-\alpha-1} \quad \text{for} \quad j = 1, 2, \ldots. \tag{3.2}$$

This condition is assumed such that we can possibly obtain a reasonable estimate of the corresponding eigenfunction of θ_j. Our adaptivity result also requires a condition on X. The process X is assumed to be left continuous (or right-continuous) at each point and that, for each $k > 0$ and some $\epsilon > 0$,

$$\sup_t \mathbb{E}\left\{|X(t)|^k\right\} < M_k \text{ and } \sup_{s,t} \mathbb{E}\left\{|s-t|^{-\epsilon}|X(t)-X(s)|^k\right\} < M_{k,\epsilon} \quad (3.3)$$

and for each $r \geq 1$,

$$\sup_{j\geq 1} \theta_j^{-r} \mathbb{E}\left\{\int (X - \mathbb{E}X)\,\phi_j\right\}^{2r} \leq M_r' \quad (3.4)$$

for some constant $M_r' > 0$.

Let $\mathcal{F}(\alpha, \beta, M)$ denote the set of distributions F of (X, Y) that satisfies (3.1) – (3.4) with $M = \{M_*, M_0, M_k, M_{k,\epsilon}, M_r'\}$. The minimax rate of convergence for estimating the slope function b over these smoothness classes has been derived by Hall and Horowitz (2007). It is shown that the minimax risk satisfies

$$\inf_{\hat{b}} \sup_{\mathcal{F}(\alpha,\beta,M)} \mathbb{E}\|\hat{b} - b\|_2^2 \asymp n^{-(2\beta-1)/(\alpha+2\beta)}. \quad (3.5)$$

The rate-optimal procedure given in Hall and Horowitz (2007) is based on frequency cut-off. Their estimator is not adaptive; it requires the knowledge of α and β. The following result shows that the block thresholding estimator $\hat{b}$ given in (2.14) is rate optimally adaptive over the collection of parameter spaces.

Theorem 1. *Under the conditions* (3.1) – (3.4) *the block thresholding estimator* $\hat{b}$ *given in* (2.14) *satisfies, for all* $2 < \alpha < \beta$,

$$\sup_{\mathcal{F}(\alpha,\beta,M)} \mathbb{E}\|\hat{b} - b\|_2^2 \leq D n^{-(2\beta-1)/(\alpha+2\beta)} \quad (3.6)$$

for some constant $D > 0$.

In addition to the function classes defined in (3.1), one can also consider adaptivity of the estimator $\hat{b}$ over other function classes. For example, consider the function classes with a Sobolev-type constraint:

$$\mathcal{S}^\beta(M_*) = \left\{ b = \sum_{j=1}^\infty b_j\,\phi_j, \text{ with } \sum_{j=1}^\infty j^{2\beta-1} b_j^2 \leq M_* \text{ for } j = 1, 2, \ldots \right\}.$$

Let $\mathcal{F}_1(\alpha, \beta, M)$ denote the set of distributions of (X, Y) that satisfies (3.2) – (3.4) and $b \in \mathcal{S}^\beta(M_*)$.

Theorem 2. *Under assumptions* (3.2) – (3.4), *the estimator* $\hat{b}$ *given in* (2.14) *satisfies, for all* $2 < \alpha < \beta$,

$$\sup_{\mathcal{F}_1(\alpha,\beta,M)} \mathbb{E}\|\hat{b} - b\|_2^2 \le Dn^{-(2\beta-1)/(\alpha+2\beta)}. \tag{3.7}$$

for some constant $D > 0$.

The proof of Theorem 2 is similar to the one for Theorem 1 with some minor modifications.

Remark 3. Theorems 1 and 2 remain true if the shrinkage factor ρ_j in (2.14) is replaced by $\rho_j = (1 - \lambda L_j \sigma^2 / n S_i^2)_+$ for any constant $\lambda > 1$.

Remark 4. We have so far focused on block thresholding. A simpler term-by-term thresholding rule can be used to yield a slightly weaker result. Let $\tilde{b}_j = \hat{\theta}_j^{-1} \hat{g}_j$ as in (2.9). Set

$$\hat{b}_j = \begin{cases} \operatorname{sgn}(\tilde{b}_j) \left(|\tilde{b}_j| - \sigma \sqrt{\dfrac{2 \log n}{n \hat{\theta}_j}} \right)_+ & \text{for } 1 \le j \le \hat{m}^*, \\ 0 & \text{for } j > \hat{m}^*. \end{cases} \tag{3.8}$$

Note that this estimator is equivalent to setting

$$\hat{d}_j = \begin{cases} \operatorname{sgn}(\tilde{d}_j) \left(|\tilde{b}_j| - \sigma \sqrt{\dfrac{2 \log n}{n}} \right)_+ & \text{for } 1 \le j \le \hat{m}^*, \\ 0 & \text{for } j > \hat{m}^*. \end{cases} \tag{3.9}$$

and $\hat{b}_j = \hat{\theta}_j^{-1/2} \hat{d}_j$. Now let $\hat{b}_t(u) = \sum_{j=1}^{\hat{m}^*} \hat{b}_j \hat{\phi}_j(u)$ with $\hat{b}_j$ given in (3.8). Then under the conditions of Theorem 1, we have

$$\sup_{\mathcal{F}(\alpha,\beta,M)} \mathbb{E}\|\hat{b}_t - b\|_2^2 \le C \left(\frac{\log n}{n} \right)^{(2\beta-1)/(\alpha+2\beta)} \tag{3.10}$$

for some constant $C > 0$. In other words, the term-by-term thresholding estimator $\hat{b}_t$ achieves the rate of convergence with a logarithmic factor of the minimax risk simultaneously over a collection of function classes. The same result holds with $\mathcal{F}(\alpha, \beta, M)$ replaced by $\mathcal{F}_1(\alpha, \beta, M)$ in (3.10).

4. Numerical Properties

The block thresholding procedure proposed in Section 2.2 is easy to implement. In this section, we investigate its numerical performance through a simulation study in two settings. In particular, we compare its finite-sample properties with those of the non-adaptive procedure introduced in Hall and Horowitz (2007).

In the first setting, the predictor X_i's were observed continuously, and independently distributed as

Table 1. Comparison of MISE in the continuous X case.

n	σ	α	β	MISE($\hat{b}$)	MISE($\hat{b}_{H,m}$)							
					$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
100	1	1.1	2	0.032	0.113	0.053	0.038	0.033	0.031	0.032	0.034	0.038
			2.5	0.027	0.080	0.039	0.029	0.028	0.028	0.030	0.034	0.038
		1.5	2	0.024	0.097	0.034	0.023	0.021	0.024	0.032	0.046	0.059
			2.5	0.016	0.055	0.019	0.015	0.017	0.025	0.033	0.044	0.058
		2	2	0.027	0.092	0.029	0.019	0.023	0.039	0.063	0.097	0.133
			2.5	0.014	0.044	0.014	0.013	0.024	0.038	0.060	0.093	0.127
	2	1.1	2	0.046	0.149	0.058	0.047	0.049	0.059	0.077	0.097	0.118
			2.5	0.038	0.080	0.041	0.041	0.047	0.060	0.074	0.094	0.113
		1.5	2	0.042	0.102	0.045	0.037	0.056	0.076	0.114	0.170	0.220
			2.5	0.029	0.059	0.028	0.033	0.050	0.073	0.108	0.160	0.212
		2	2	0.042	0.087	0.036	0.045	0.069	0.129	0.227	0.354	0.500
			2.5	0.028	0.046	0.021	0.038	0.088	0.145	0.226	0.340	0.552
500	1	1.1	2	0.007	0.091	0.027	0.014	0.009	0.008	0.007	0.007	0.008
			2.5	0.005	0.045	0.012	0.007	0.006	0.005	0.006	0.006	0.007
		1.5	2	0.007	0.084	0.022	0.010	0.007	0.007	0.008	0.009	0.012
			2.5	0.004	0.038	0.008	0.005	0.004	0.004	0.006	0.007	0.104
		2	2	0.010	0.087	0.022	0.010	0.008	0.010	0.012	0.016	0.024
			2.5	0.005	0.038	0.007	0.004	0.005	0.008	0.013	0.018	0.024
	2	1.1	2	0.014	0.093	0.028	0.016	0.014	0.015	0.017	0.021	0.026
			2.5	0.009	0.045	0.013	0.009	0.010	0.011	0.013	0.016	0.021
		1.5	2	0.014	0.089	0.024	0.014	0.013	0.018	0.023	0.035	0.043
			2.5	0.008	0.042	0.010	0.008	0.010	0.014	0.021	0.030	0.043
		2	2	0.017	0.083	0.023	0.015	0.017	0.025	0.043	0.079	0.118
			2.5	0.009	0.041	0.009	0.009	0.015	0.026	0.041	0.068	0.097

$$X = \sum_j \gamma_j W_j \phi_j(t), \quad \text{for} \quad t \in [0,2],$$

where $\gamma_j = (-1)^{j+1} j^{-\alpha/2}$, $\{\phi_j\}_{j=1}^{\infty} = \{1, \sin(\pi t), \cos(\pi t), \sin(2\pi t), \cos(2\pi t), \dots\}$ is the Fourier series with period 2, and W_j's are i.i.d. standard normal variables. In addition, the slope function b was taken to be $b(t) = \sum_j j^{-\beta}\phi_j(t)$ and the errors Z_i were distributed as normal $N(0, \sigma^2)$. To show the advantage of the adaptive procedure, we compared our estimator $\hat{b}$ with that of Hall and Horowitz (2007), denoted by $\hat{b}_{H,m}$, which was shown to be minimax optimal but not adaptive. In the construction of their estimator $\hat{b}_{H,m}$, one needs to specify the optimal cut-off index m, which requires the knowledge of the smoothing parameters that are usually unknown in practice. We compared the MISE (1.2) between our adaptive procedure and their method with different values of m chosen from $\{1, 2, \dots, 8\}$. Similar to the setting in Hall and Horowitz (2007), we took a range of values for $(\sigma, n, \alpha, \beta)$ in our simulation study. Specifically,

Table 2. Comparison of MISE in the discrete X case.

n	σ	α	β	MISE($\hat{b}$)	MISE($\hat{b}_{H,m}$)							
					$m=1$	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
100	1	1.1	2	0.036	0.131	0.059	0.042	0.036	0.035	0.035	0.038	0.043
			2.5	0.030	0.088	0.040	0.031	0.031	0.033	0.034	0.038	0.042
		1.5	2	0.032	0.097	0.040	0.031	0.029	0.033	0.042	0.053	0.063
			2.5	0.021	0.057	0.024	0.020	0.021	0.028	0.037	0.045	0.055
		2	2	0.030	0.101	0.034	0.026	0.028	0.036	0.048	0.065	0.086
			2.5	0.017	0.048	0.017	0.018	0.028	0.039	0.052	0.074	0.100
	2	1.1	2	0.052	0.124	0.064	0.053	0.054	0.064	0.075	0.087	0.108
			2.5	0.044	0.091	0.050	0.043	0.047	0.054	0.068	0.086	0.111
		1.5	2	0.048	0.099	0.046	0.043	0.056	0.077	0.109	0.139	0.187
			2.5	0.036	0.061	0.032	0.039	0.050	0.070	0.096	0.133	0.173
		2	2	0.046	0.102	0.042	0.046	0.074	0.116	0.189	0.257	0.350
			2.5	0.031	0.053	0.029	0.043	0.072	0.122	0.180	0.250	0.335
500	1	1.1	2	0.008	0.094	0.027	0.014	0.010	0.009	0.008	0.009	0.010
			2.5	0.006	0.044	0.013	0.008	0.007	0.006	0.007	0.007	0.008
		1.5	2	0.009	0.090	0.026	0.013	0.009	0.009	0.009	0.010	0.012
			2.5	0.005	0.043	0.010	0.006	0.005	0.006	0.007	0.009	0.011
		2	2	0.012	0.086	0.023	0.011	0.010	0.010	0.013	0.018	0.022
			2.5	0.006	0.040	0.009	0.005	0.006	0.009	0.012	0.017	0.022
	2	1.1	2	0.015	0.090	0.028	0.017	0.014	0.015	0.017	0.020	0.024
			2.5	0.010	0.047	0.014	0.010	0.010	0.012	0.015	0.019	0.023
		1.5	2	0.016	0.089	0.025	0.016	0.014	0.017	0.022	0.029	0.037
			2.5	0.009	0.043	0.011	0.009	0.011	0.017	0.024	0.032	0.040
		2	2	0.018	0.085	0.025	0.017	0.019	0.029	0.041	0.058	0.076
			2.5	0.010	0.041	0.010	0.010	0.014	0.021	0.031	0.046	0.063

$(\sigma, n, \alpha, \beta)$ was chosen from the set $\{1,2\} \times \{100, 500\} \times \{1.1, 1.5, 2.0\} \times \{2, 2.5\}$. All the results were based on 200 Monte Carlo replications for each parameter setting and the MISEs of different procedures are recorded in Table 1.

The choices of X, b and parameters $(\sigma, n, \alpha, \beta)$ in the second setting are the same as those for the first, except that each X_i was observe discretely on an equally spaced grid of 41 points on $[0, 2]$, with i.i.d. additive $N(0, 4)$ random noise. We used a Fourier basis smoother to estimate functions X_i's from the discrete data. Table 2 summarizes the averaged MISEs of the proposed block thresholding method and the method of Hall and Horowitz (2007) with different cut-off index m, computed by averaging over 200 Monte Carlo simulations. It is clear from these results that both procedures are robust against discretization and random errors.

The results in both Table 1 and Table 2 show that the MISEs of the method proposed by Hall and Horowitz (2007) are sensitive to the choice of cut-off index

m. The optimal choice of m is usually unknown in practice and needs to be tuned via empirical method such as cross validation. In contrast, our proposed method is purely data-driven and adaptive to different degrees of smoothness. Both tables demonstrate the advantage of the proposed adaptive procedure. The performance of the proposed adaptive method is as good as, if not better than, the method of Hall and Horowitz (2007) with the optimally chosen m.

Supplementary Materials

Supplement Cai, Zhang and Zhou (2017) includes the proofs of the main results as well as some technical lemmas.

Acknowledgment

The research of Tony Cai was supported in part by NSF Grant DMS-1403708 and NIH Grant R01 CA127334. The research of Harrison Zhou was supported in part by NSF Grant DMS-1507511.

References

Cai, T. T. (1999). Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Ann. Statist.* **27**, 898–924.

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

Cai, T. T., Low, M. and Zhao, L. (2009). Sharp adaptive estimation by a blockwise method. *J. Nonparametr. Stat.* **21**, 839–850.

Cai, T. T., Zhang, L. and Zhou, H.H. (2017). Supplement to "Adaptive functional linear regression via functional principal component analysis and block thresholding". `http://www3.stat.sinica.edu.tw/statistica/`

Cardot and Sarda, P. (2006). Linear regression models for functional data. In *The Art of Semiparametrics*, 49–66. Physica-Verlag HD.

Cuevas, A., Febrero, M. and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response. *Canad. J. Statist.* **30**, 285–300.

Efromovich, S. Y. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30**, 557–661.

Ferraty, F. and Vieu, P. (2000). Fractal dimensionality and regression estimation in semi-normed vectorial spaces. *C. R. Acad. Sci. Paris Sér. I* **330**, 139–142.

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

Hall, P. and Horowitz, J. L. (2007). Methodology and convergence 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. R. Stat. Soc. Ser. B. Stat. Methodol.* **68**, 109–126.

Hall, P., Kerkyacharian, G. and Picard, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26**, 922–942.

Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Anal.* **98**, 1782–1804.

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

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis,* 2nd Edition. Springer, New York.

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: tcai@wharton.upenn.edu

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: linjunz@wharton.upenn.edu

Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA.

E-mail: huibin.zhou@yale.edu

(Received February 2017; accepted September 2017)