Adaptive Functional Linear Regression

T. Tony Cai[1] and Harrison H. Zhou[2]
University of Pennsylvania and 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.

Keywords: Adaptive estimation; Block thresholding; Eigenfunction; Eigenvalue; Functional data analysis; Minimax estimation; Principal components analysis; Rate of convergence; Slope function; Smoothing; Spectral decomposition.
AMS 2000 Subject Classification: Primary 62J05; secondary 62G20.

[1]Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104.
The research of Tony Cai was supported in part by NSF Grant DMS-0604954.
[2]Department of Statistics, Yale University, New Haven, CT 06511. The research of Harrison Zhou was supported in part by NSF Grants DMS-0645676.

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

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

where the response Y_i and the intercept a are scalar, 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 $\{(X_i, Y_i) : i = 1, ...n\}$. 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)

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

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 *et al.* (2002); Cardot and Sarda (2003); 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 (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 and 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 (2000) 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 (2000) 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 paper is organized as follows. In Section 2, after basic notation 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. The proofs are given in Section 4.

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 components 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 $\{(X_i, Y_i) : i = 1, ...n\}$ as in (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

$$\begin{aligned} K(u,v) &= \mathrm{cov}\{X(u), X(v)\} \\ \hat{K}(u,v) &= \frac{1}{n}\sum_{i=1}^{n}\{X_i(u) - \bar{X}(u)\}\{X_i(v) - \bar{X}(v)\} \end{aligned}$$

where $\bar{X} = \frac{1}{n}\sum X_i$. 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

$$K(u,v) = \sum_{j=1}^{\infty}\theta_j\phi_j(u)\phi_j(v), \quad \hat{K}(u,v) = \sum_{j=1}^{\infty}\hat{\theta}_j\hat{\phi}_j(u)\hat{\phi}_j(v), \tag{3}$$

where

$$\theta_1 > \theta_2 > ... > 0, \text{ and } \hat{\theta}_1 \geq \hat{\theta}_2 \geq ... \geq \hat{\theta}_{n+1} = \ldots = 0 \tag{4}$$

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) can be rewritten as

$$Y_i = \mu + \int \left[X_i - \mathbb{E}(X)\right] b + Z_i,\ i = 1, 2, ..., n \tag{5}$$

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

$$X_i - \mathbb{E}X = \sum_{j=1}^{\infty} x_{i,j}\phi_j \tag{6}$$

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 (5) can be written as

$$Y_i = \mu + \sum_{j=1}^{\infty} x_{i,j}b_j + Z_i,\ i = 1, 2, ..., n \tag{7}$$

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 (7) 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

$$X_i - \bar{X} = \sum_{j=1}^{n} \hat{x}_{i,j}\hat{\phi}_j \quad \text{for } i = 1, 2, ..n \tag{8}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i,j}\hat{x}_{i,k} = \int\int \hat{K}(u,v)\hat{\phi}_j(u)\hat{\phi}_k(v) = \hat{\theta}_j\delta_{j,k} \tag{9}$$

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

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

Hence

$$Y_i - \bar{Y} = \sum_{j=1}^{n} \hat{x}_{i,j}\check{b}_j + Z_i - \bar{Z},\ i = 1, 2, ..., n \tag{10}$$

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 equation (10) 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 equation (10) are orthogonal to each other and $\sum_{i=1}^{n} \hat{x}_{i,j}^2 = \hat{\theta}_j n$ from equation (9), for $\hat{\theta}_j \neq 0$ we may estimate $\check{b}_j$ (or b_j) by

$$\begin{aligned} \tilde{b}_j &= \hat{\theta}_j^{-1} n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})\hat{x}_{i,j} = \hat{\theta}_j^{-1} n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \int \left[X_i(u) - \bar{X}(u)\right] \hat{\phi}_j(u) \qquad (11) \\ &= \hat{\theta}_j^{-1} \int \hat{g}(u)\hat{\phi}_j(u) = \hat{\theta}_j^{-1}\hat{g}_j \end{aligned}$$

where

$$\hat{g}(u) = n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y}) \left[X_i(u) - \bar{X}(u)\right] \text{ and } \hat{g}_j = \int \hat{g}\hat{\phi}_j. \tag{12}$$

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

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

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 equation (11) 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 (2000) 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 4 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 the notation introduced above the block thresholding procedure can then be described in detail as follows. Let

$$\hat{m}^* = \arg\min\left\{m : \hat{\theta}_m/\hat{\theta}_1 \le n^{-1/3}\right\}. \tag{13}$$

It will be shown in Section 4 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

$$\tilde{d}_j = \hat{\theta}_j^{-\frac{1}{2}}\tilde{g}_j \quad \text{and} \quad d_j = \theta_j^{-\frac{1}{2}}g_j. \tag{14}$$

Lemma 4 in Section 4 shows that the variance $\mathbb{V}ar(\tilde{d}_j) = \frac{\sigma^2}{n}(1+o(1))$ and so $\tilde{d}_j$ 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 $\hat{b}_j = \hat{\theta}_j^{-\frac{1}{2}}\hat{d}_j$.

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

1. Divide the indices $\{1, 2, ..., \hat{m}^*\}$ into nonoverlapping blocks B_1, B_2, ..., 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

$$\hat{d}_j = (1 - \frac{2L_i\sigma^2}{nS_i^2})_+ \cdot \tilde{d}_j \tag{15}$$

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

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

$$\hat{b}(u) = \sum_{j=1}^{\hat{m}^*}\hat{b}_j\hat{\phi}_j(u) = \sum_{j=1}^{\hat{m}^*}\rho_j\tilde{b}_j\hat{\phi}_j(u) \tag{16}$$

   where $\rho_j = (1 - \frac{2L_j\sigma^2}{nS_i^2})_+$ 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 (15) because of its simplicity. In addition to the James-Stein rule, other shrinkage rules such as the blockwise hard thresholding rule*

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

can be used as well.

Remark 2 *In the procedure we assume σ is known, since it can be estimated easily. In equation (10), we may apply principal components regression by replacing "n" in equation (10) with a constant $m = \log^2 n$. Let $\hat{\mathrm{b}}_j$ be the ordinary least squares estimate of $\check{\mathrm{b}}_j$. 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 property

We now turn to the asymptotic properties of the block thresholding procedure for the functional linear regression under the mean integrated squared error (2). The theoretical results show that the block thresholding estimator given in (16) 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

$$\mathcal{B}^{\beta}(M_*) = \left\{ b = \sum_{j=1}^{\infty} b_j \, \phi_j, \text{ with } |b_j| \leq M_* \, j^{-\beta} \text{ for } j = 1, 2, ... \right\}. \tag{17}$$

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

$$M_0^{-1} j^{-\alpha} \leq \theta_j \leq M_0 j^{-\alpha}, \quad \theta_j - \theta_{j+1} \geq M_0^{-1} j^{-\alpha-1} \quad \text{for} \quad j = 1, 2, ... \tag{18}$$

This condition is assumed such that we may possibly obtain a reasonable estimate of the corresponding eigenfunction of θ_j. Our adaptivity result also requires the following 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 $\epsilon > 0$

$$\sup_t \mathbb{E}\left\{|X(t)|^k\right\} < M_k \text{ and } \sup_{s,t} \mathbb{E}\left\{|s-t|^{-\epsilon} |X(t) - X(s)|^k\right\} < M_{k,\epsilon} \tag{19}$$

and for each $r \geq 1$,

$$\sup_{j \geq 1} \theta_j^{-r} \mathbb{E}\left[\int (X - \mathbb{E}X)\, \phi_j\right]^{2r} \leq M_r' \tag{20}$$

for some constant $M_r' > 0$.

Let $\mathcal{F}(\alpha, \beta, M)$ denote the set of distributions F of (X, Y) that satisfies (17) - (20) 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

$$\inf_{\hat{b}} \sup_{\mathcal{F}(\alpha,\beta,M)} \mathbb{E}\|\hat{b} - b\|_2^2 \asymp n^{-\frac{2\beta-1}{\alpha+2\beta}}. \tag{21}$$

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 (16) is rate optimally adaptive over the collection of parameter spaces.

Theorem 1 *Under the conditions (17) - (20) the block thresholding estimator $\hat{b}$ given in (16) satisfies, for all $2 < \alpha < \beta$,*

$$\sup_{\mathcal{F}(\alpha,\beta,M)} \mathbb{E}\|\hat{b} - b\|_2^2 \leq D n^{-\frac{2\beta-1}{\alpha+2\beta}} \tag{22}$$

for some constant $D > 0$.

In addition to the function classes defined in (17), one can also consider adaptivity of the estimator $\hat{b}$ over other function classes. For example, consider the following function classes with a Sobolev-type constraint:

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

Let $\mathcal{F}_1(\alpha, \beta, M)$ denote the set of distributions of (X, Y) that satisfies (18) - (20) and $b \in \mathcal{S}^\beta(M_*)$.

Theorem 2 *Under assumptions (18) - (20), the estimator $\hat{b}$ given in (16) satisfies, for all $2 < \alpha < \beta$,*

$$\sup_{\mathcal{F}_1(\alpha,\beta,M)} \mathbb{E}\|\hat{b} - b\|_2^2 \leq Dn^{-\frac{2\beta-1}{\alpha+2\beta}}. \tag{23}$$

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 (16) is replaced by $\rho_j = (1 - \frac{\lambda L_j \sigma^2}{n S_i^2})_+$ 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 (11). Set*

$$\hat{b}_j = \begin{cases} \operatorname{sgn}(\tilde{b}_j)(|\tilde{b}_j| - \sigma\sqrt{\frac{2\log n}{n\hat{\theta}_j}})_+ & \text{for } 1 \leq j \leq \hat{m}^* \\ 0 & \text{for } j > \hat{m}^* \end{cases}. \tag{24}$$

Note that this estimator is equivalent to setting

$$\hat{d}_j = \begin{cases} \operatorname{sgn}(\tilde{d}_j)(|\tilde{b}_j| - \sigma\sqrt{\frac{2\log n}{n}})_+ & \text{for } 1 \leq j \leq \hat{m}^* \\ 0 & \text{for } j > \hat{m}^* \end{cases} \tag{25}$$

and $\hat{b}_j = \hat{\theta}_j^{-\frac{1}{2}} \hat{d}_j$. Now let $\hat{b}_t(u) = \sum_{j=1}^{\hat{m}^*} \hat{b}_j \hat{\phi}_j(u)$ with $\hat{b}_j$ given in (24). Then under the conditions of Theorem 1, we have

$$\sup_{\mathcal{F}(\alpha,\beta,M)} \mathbb{E}\|\hat{b}_t - b\|_2^2 \leq C\left(\frac{\log n}{n}\right)^{\frac{2\beta-1}{\alpha+2\beta}} \tag{26}$$

for some constant $C > 0$. In other words, the term-by-term thresholding estimator $\hat{b}_t$ is simultaneously with a logarithmic factor of the minimax risk 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 (26).

4 Proofs

We shall only prove Theorem 1. The proof of Theorem 2 is similar and thus omitted. Before we present the proof of the main result, we first collect a few technical lemmas. These auxiliary lemmas will be proved in Section 4.3. We sharpen some results in Hall and Horowitz (2007) and give a risk bound for block James-Stein estimator. In this section we shall denote by C a generic constant which may vary from place to place.

4.1 Technical lemmas

It was proposed in Hall and Horowitz (2007) to estimate b by $\sum_{j=1}^{m} \tilde{b}_j \hat{\phi}_j$ with a choice of cutoff $m = n^{\frac{1}{\alpha+2\beta}}$ to obtain minimax rate of convergence. The lemma below explains why there is no need ever to go beyond the $\hat{m}^*$-th term in defining the block thresholding procedure (16).

Lemma 1 *Let γ and γ_1 be constants satisfying $\frac{1}{\alpha+2\beta} < \gamma < \frac{1}{3\alpha} < \gamma_1$ For all $D > 0$, there exists a constant C_D such that*

$$\mathbb{P}\left(n^{\gamma} \leq \hat{m}^* \leq n^{\gamma_1}\right) \geq 1 - c_D n^{-D}$$

where $\hat{m}^*$ is defined in (13).

In this section we set

$$\frac{1}{\alpha+2\beta} < \gamma < \min\left\{\frac{1+\varepsilon}{\alpha+2\beta}, \frac{1}{3\alpha}\right\}, \frac{1}{3\alpha} < \gamma_1 < \frac{1}{2\left(\alpha+1\right)} \tag{27}$$

for a small $0 < \varepsilon < \min\left\{\frac{\alpha-2}{3}, \frac{2\beta-\alpha}{3\alpha+1}\right\}$. We give upper bounds to approximate eigenfunction ϕ_j by empirical eigenfunction $\hat{\phi}_j$ for $j \leq n^{\gamma_1}$.

Lemma 2 *For all $j \leq n^{\gamma_1}$, we have*

$$n\mathbb{E}\left\|\hat{\phi}_j - \phi_j\right\|^2 \leq Cj^2$$

and for any given $0 < \delta < 1$ and for all $D > 0$ there exists a constant $C_D > 0$ such that

$$\mathbb{P}\left\{n^{1-\delta}\left\|\hat{\phi}_j - \phi_j\right\|^2 \geq Cj^2\right\} \leq C_D n^{-D}.$$

Lemma 3 gives a variance bound for $\check{b}_j$, which helps us show that the variance of $\tilde{d}_j$ is approximately $\frac{\sigma^2}{n}$. This result is crucial for proposing a practical block thresholding procedure.

Lemma 3 *For $j \leq n^{\gamma_1}$ with $\gamma_1 < \frac{1}{2(\alpha+1)}$,*

$$\mathbb{E}\left(\check{b}_j - b_j\right)^2 \leq Cj^2/n.$$

In particular, this implies $\mathbb{V}ar(\check{b}_j) \leq Cj^2/n$ and $\mathbb{V}ar(\tilde{b}_j) = \sigma^2\theta_j^{-1}n^{-1}\left(1+o\left(1\right)\right)$.

The following lemma gives bounds for the variance and mean squared error of $\tilde{d}_j$.

Lemma 4 *For $j \leq n^{\gamma_1}$ with $\gamma_1 < \frac{1}{2(\alpha+1)}$,*

$$\mathbb{V}ar(\tilde{d}_j) = \frac{\sigma^2}{n}\left(1+o\left(1\right)\right) \quad \text{and} \quad \mathbb{E}\left(\tilde{d}_j - \theta_j^{\frac{1}{2}}b_j\right)^2 \leq Cn^{-1}j^{2-\alpha}.$$

The following two lemmas will be used to analyze the factor ρ_j in equation (16).

Lemma 5 *Let $n^{\gamma} \leq m_1 \leq m_2 \leq n^{\gamma_1}$ and $m_2 - m_1 \geq n^{\delta}$ for some $\delta > 0$. Define $S^2 = \sum_{j=m_1}^{m_2}\tilde{d}_j^2$. For any given $\varepsilon > 0$ and all $D > 0$ there exists a constant $C_D > 0$ such that*

$$\mathbb{P}(S^2 > (1+\varepsilon)\left(m_2 - m_1\right)\frac{\sigma^2}{n}) \leq C_D n^{-D}.$$

Lemma 6 *Let $\tilde{d}_j = d_j' + \epsilon_j$ where $d_j' = E(\tilde{d}_j)$. Let $\varepsilon > 0$ be a fixed constant. If the block size $L_i = \mathrm{Card}(B_i) \geq n^{\delta}$ for some $\delta > 0$, then for any $D > 0$, there exists a constant $C_D > 0$ such that*

$$\mathbb{P}(\sum_{j\in B_i}\epsilon_j^2 > (1+\varepsilon)\, L_i\frac{\sigma^2}{n}) \leq C_D n^{-D}. \tag{28}$$

And for all blocks B_i,

$$\mathbb{E}\sum_{j\in B_i}\epsilon_j^2 \leq CL_i\frac{\sigma^2}{n}. \tag{29}$$

Conventional oracle inequalities were derived for Gaussian errors. In the current setting the errors are non-Gaussian. The following lemma gives an oracle inequality for a block thresholding estimator in the case of general error distributions. See Brown, Cai, Zhang, Zhao and Zhou (2007) for a proof.

Lemma 7 *Suppose $y_i = \theta_i + \epsilon_i, \quad i = 1, ..., L$, where θ_i are constants and Z_i are random variables. Let $S^2 = \sum_{i=1}^{L} y_i^2$ and let*

$$\hat{\theta}_i = (1 - \frac{\lambda L}{S^2})_+ y_i.$$

Then

$$\mathbb{E}\|\hat{\theta} - \theta\|_2^2 \leq \min\{\|\theta\|_2^2, 4\lambda L\} + 4\mathbb{E}\|\epsilon\|_2^2 I(\|\epsilon\|_2^2 > \lambda L). \tag{30}$$

4.2 Proof of Theorem 1

We shall prove Theorem 1 for a general block thresholding estimator with the shrinkage factor $\rho_j = (1 - \frac{\lambda L_j \sigma^2}{n S_i^2})_+$ for a constant $\lambda > 1$.

Let γ and γ_1 be constants satisfying

$$\frac{1}{\alpha + 2\beta} < \gamma < \min\left\{\frac{1+\varepsilon}{\alpha+2\beta}, \frac{1}{3\alpha}\right\} \le \frac{1}{3\alpha} < \gamma_1 < \frac{1}{2(\alpha+1)}$$

for a small $\varepsilon > 0$. Let $m_* = n^\gamma$ and write $\hat{b}$ as

$$\hat{b}(u) = \sum_{j=1}^{m_*} \rho_j \tilde{b}_j \hat{\phi}_j(u) + \sum_{j=m_*+1}^{n} \rho_j \tilde{b}_j \hat{\phi}_j(u). \tag{31}$$

We shall show that $\mathbb{E}\|\hat{b} - b\|_2^2 \le C n^{-\frac{2\beta-1}{\alpha+2\beta}}$. Note that

$$\begin{aligned}
\mathbb{E}\|\hat{b} - b\|_2^2 &= \mathbb{E}\|\sum_{j=1}^{m_*} \hat{b}_j\hat{\phi}_j(u) + \sum_{j=m_*+1}^{n} \hat{b}_j\hat{\phi}_j(u) - \sum_{j=1}^{m_*} b_j\phi_j(u) - \sum_{j=m_*+1}^{\infty} b_j\phi_j(u)\|_2^2 \\
&\le 3\mathbb{E}\|\sum_{j=1}^{m_*} \hat{b}_j\hat{\phi}_j(u) - \sum_{j=1}^{m_*} b_j\phi_j(u)\|_2^2 + 3\sum_{j=m_*+1}^{n} \mathbb{E}(\hat{b}_j^2) + 3\sum_{j=m_*+1}^{\infty} b_j^2. \qquad (32)
\end{aligned}$$

The last term (32) is bounded by $Cn^{-\gamma(2\beta-1)} = o\left(n^{-(2\beta-1)/(\alpha+2\beta)}\right)$ since $\gamma > \frac{1}{\alpha+2\beta}$. We first show that the second term (32) is small as well. Let $m^* = n^{\gamma_1}$ and let i_* and i^* be the corresponding block indices of the (m_*+1)-st and m^*-th term respectively. (That is, b_{m_*+1} is in the i_*-th block and b_{m^*} is in the i^*-th block.) Then it follows from Lemmas 1 and 5 that

$$\begin{aligned}
\sum_{j=m_*+1}^{n} \mathbb{E}(\hat{b}_j^2) &= \left(\sum_{j=m_*+1}^{m^*} + \sum_{j=m^*+1}^{n}\right) \mathbb{E}(\rho_j^2 \tilde{b}_j^2) \\
&\le \sum_{j=m_*+1}^{m^*} (\mathbb{E}\rho_j^4)^{\frac{1}{2}} (\mathbb{E}\tilde{b}_j^4)^{\frac{1}{2}} + \sum_{j=m^*+1}^{n} (\mathbb{E}\tilde{b}_j^4)^{\frac{1}{2}} \mathbb{P}^{\frac{1}{2}}\left(\hat{m}^* \ge n^{\gamma_1} + 1\right) \\
&\le \sum_{i=i_*}^{i^*} \left[\mathbb{P}(S_i^2 > \lambda L \sigma^2/n)\right]^{1/2} \sum_{j \in B_i} (\mathbb{E}\tilde{b}_j^4)^{\frac{1}{2}} + \sum_{j=m^*+1}^{n} (\mathbb{E}\tilde{b}_j^4)^{\frac{1}{2}} \left[\mathbb{P}\left(\hat{m}^* \ge n^{\gamma_1} + 1\right)\right]^{1/2} \\
&= o\left(n^{-\frac{2\beta-1}{\alpha+2\beta}}\right).
\end{aligned}$$

We now turn to the first and dominant term in (32). The Cauchy-Schwarz inequality yields

$$\begin{aligned}
\mathbb{E}\|\sum_{j=1}^{m_*} \hat{b}_j\hat{\phi}_j(u) - \sum_{j=1}^{m_*} b_j\phi_j(u)\|_2^2 &\le 2\mathbb{E}\|\sum_{j=1}^{m_*} (\hat{b}_j - b_j)\hat{\phi}_j(u)\|_2^2 + 2\mathbb{E}\|\sum_{j=1}^{m_*} b_j(\hat{\phi}_j(u) - \phi_j(u))\|_2^2 \\
&\le 2\sum_{j=1}^{m_*} \mathbb{E}(\hat{b}_j - b_j)^2 + 2m_* \sum_{j=1}^{m_*} b_j^2 \mathbb{E}\|\hat{\phi}_j(u) - \phi_j(u)\|_2^2.
\end{aligned}$$

Lemma 2 implies the second term in the equation above is bounded by

$$C\frac{m_*}{n}\sum_{j=1}^{m_*} b_j^2 j^2 = O\left(n^{\gamma-1}\right) = o\left(n^{-\frac{2\beta-1}{\alpha+2\beta}}\right)$$

since $\sum_{j=1}^{m_*} b_j^2 j^2$ is finite and $\gamma < \frac{\alpha+1}{\alpha+2\beta}$ which implies $\gamma - 1 < -\frac{2\beta-1}{\alpha+2\beta}$. Set $d'_j = \mathbb{E}(\tilde{d}_j)$. Let κ_i be the smallest eigenvalue in the B_i-th block. Then

$$\begin{aligned}
&\sum_{j=1}^{m_*}\mathbb{E}(\hat{b}_j - b_j)^2 = \sum_{j=1}^{m_*}\mathbb{E}(\hat{\theta}_j^{-\frac{1}{2}}\hat{d}_j - \theta_j^{-\frac{1}{2}}d_j)^2 \le 2\sum_{j=1}^{m_*}\theta_j^{-1}\mathbb{E}(\hat{d}_j - d_j)^2 + 2\sum_{j=1}^{m_*}\mathbb{E}\left[\hat{d}_j^2(\hat{\theta}_j^{-\frac{1}{2}} - \theta_j^{-\frac{1}{2}})^2\right] \\
\le\ & 2\sum_{j=1}^{m_*}\theta_j^{-1}\mathbb{E}(\hat{d}_j - d_j)^2 + 2\sum_{j=1}^{m_*}\mathbb{E}\left[\tilde{d}_j^2(\hat{\theta}_j^{-\frac{1}{2}} - \theta_j^{-\frac{1}{2}})^2\right] \\
\le\ & 2\sum_{i=1}^{i_*}\kappa_i^{-1}\sum_{j\in B_i}\mathbb{E}(\hat{d}_j - d'_j)^2 + 2\sum_{i=1}^{i_*}\kappa_i^{-1}\sum_{j\in B_i}(d'_j - d_j)^2 + 2\sum_{j=1}^{m_*}\mathbb{E}\left[\tilde{d}_j^2(\hat{\theta}_j^{-\frac{1}{2}} - \theta_j^{-\frac{1}{2}})^2\right] \\
\equiv\ & T_1 + T_2 + T_3.
\end{aligned}$$

From equations (34) and (35) and Lemma 4, it is easy to see

$$T_3 \le C\sum_{j=1}^{m_*}\mathbb{E}\{\tilde{d}_j^2\theta_j^{-3}(\hat{\theta}_j - \theta_j)^2\} = o\left(n^{-\frac{2\beta-1}{\alpha+2\beta}}\right).$$

We now turn to the dominant term $T_1 + T_2$. This term is most closely related to the block thresholding rule and we need to show that $T_1 + T_2 \le Cn^{-\frac{2\beta-1}{\alpha+2\beta}}$. To bound T_1, it is necessary to analyze the risk of the block thresholding rule for a single block B_i. It follows from Lemma 7 that

$$\sum_{j\in B_i}\mathbb{E}(\hat{d}_j - d'_j)^2 \le \min\{4\lambda L_i\sigma^2/n, \sum_{j\in B_i}(d'_j)^2\} + 4\mathbb{E}\{(\sum_{j\in B_i}\epsilon_j^2)\cdot I(\sum_{j\in B_i}\epsilon_j^2 > \lambda L_i\sigma^2/n)\} \quad (33)$$

where $\lambda > 1$ is a constant. Lemma 4 implies

$$\left(d'_j - \theta_j^{\frac{1}{2}}b_j\right)^2 \le Cn^{-1}j^{2-\alpha}.$$

Note that for all j in B_i, we have $\theta_j^{-1} \asymp \kappa_i^{-1}$. Hence for $m_* = n^\gamma$ with $\gamma < \frac{1+\varepsilon}{\alpha+2\beta}$ we have

$$T_2 \le C\sum_{j=1}^{m_*}\theta_j^{-1}n^{-1}j^{2-\alpha} \le \frac{C_1}{n}\left(1 + m_*^3\right) = o\left(n^{-\frac{2\beta-1}{\alpha+2\beta}}\right)$$

Let $m = n^{\frac{1}{\alpha+2\beta}}$, then equation (33) and Lemma 6 give

$$T_1 \le C\sum_{j=1}^{m}\frac{j^\alpha}{n} + C\sum_{j=m+1}^{m_*}\left[\theta_j^{-1}\cdot\left(\theta_j^{1/2}b_j\right)^2 + \theta_j^{-1}n^{-1}j^{2-\alpha}\right] + C/n \le C_1 n^{-\frac{2\beta-1}{\alpha+2\beta}}.$$

These together imply $\mathbb{E}\|\hat{b} - b\|_2^2 \le Cn^{-\frac{2\beta-1}{\alpha+2\beta}}$. ∎

4.3 Proof of auxiliary lemmas

Let $\Delta^2 = \left\|\hat{K} - K\right\|^2 = \int\int \left(\hat{K}(u,v) - K(u,v)\right)^2 dudv$ and $\tau_j = \min_{k\leq j}(\theta_k - \theta_{k+1})$. It is known in Bhatia, Davis and McIntosh (1983) that

$$\sup_j \left|\hat{\theta}_j - \theta_j\right| \leq \Delta, \ \sup_{j\geq 1} \tau_j \left\|\hat{\phi}_j - \phi_j\right\| \leq 8^{1/2}\Delta. \tag{34}$$

For $\varepsilon > 0$, it was shown in Hall and Hosseini-Nasab (2006, Lemma 3.3)

$$\mathbb{P}\left(\Delta > n^{\varepsilon - 1/2}\right) = c_D n^{-D} \tag{35}$$

for each $D > 0$ under the assumption (19).

It is useful to rewrite $\tilde{b}_j$ as

$$\begin{aligned}
\tilde{b}_j &= \hat{\theta}_j^{-1}\hat{g}_j = \hat{\theta}_j^{-1}\int \frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y})\{X_i(u) - \bar{X}(u)\}\hat{\phi}_j(u) \\
&= \hat{\theta}_j^{-1}\int \frac{1}{n}\sum_{i=1}^n \left(\langle X_i - \bar{X}, b\rangle + Z_i - \bar{Z}\right)\{X_i(u) - \bar{X}(u)\}\hat{\phi}_j(u) \\
&= \check{b}_j + \hat{\theta}_j^{-1}\frac{1}{n}\int (\underline{X} - \bar{X})'\hat{\phi}_j \cdot (\underline{Z} - \bar{Z}) = \check{b}_j + \hat{\theta}_j^{-1}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{Z} - \bar{Z}).
\end{aligned}$$

Using the fact that for any two random variables X and Y, $\mathbb{V}ar(Y) = \mathbb{E}(\mathbb{V}ar(Y|X)) + \mathbb{V}ar(\mathbb{E}(Y|X))$ and the facts that $\underline{Z}$ has mean zero and is independent of $\underline{X}$, we have

$$\mathbb{V}ar(\tilde{b}_j) = \mathbb{V}ar(\check{b}_j) + \frac{\sigma^2}{n^2}\sum_{i=1}^n \mathbb{E}(\hat{\theta}_j^{-2}\hat{x}_{i,j}^2) = \mathbb{V}ar(\check{b}_j) + \frac{\sigma^2}{n}\mathbb{E}\hat{\theta}_j^{-1}.$$

4.3.1 Proof of Lemma 1

Recall that $\hat{m}^* = \arg\min\left\{m : \hat{\theta}_m/\hat{\theta}_1 \leq n^{-1/3}\right\}$. Note that $\theta_j \geq M_0^{-1} j^{-\alpha}$. Since γ satisfies $\frac{1}{\alpha+2\beta} < \gamma < \frac{1}{3\alpha}$, then for $m \leq n^\gamma$ we have $\theta_m \geq M_0^{-1} n^{-\alpha\gamma}$. Since $\alpha\gamma < 1/3$, the equations (34) and (35) imply that for any $D > 0$ there exists a constant $C_D > 0$ such that

$$\mathbb{P}\left(\cup_{m=1}^{n^\gamma}\left\{\hat{\theta}_m/\hat{\theta}_1 \leq n^{-1/3}\right\}\right) \leq c_D n^{-D}$$

and hence

$$\mathbb{P}(\hat{m}^* \leq n^\gamma) \leq c_D n^{-D}, \text{ i.e., } \mathbb{P}(\hat{m}^* \geq n^\gamma) \geq 1 - c_D n^{-D}.$$

Similarly, for $m \geq n^{\gamma_1}$ we have

$$\theta_m \leq M_0 n^{-\gamma_1\alpha}$$

with $\alpha\gamma_1 > 1/3$, then

$$\mathbb{P}\left(\cup_{n\geq m\geq n^{\gamma_1}}\left\{\hat{\theta}_m/\hat{\theta}_1 > n^{-1/3}\right\}\right) \geq c_D n^{-D}$$

and hence

$$\mathbb{P}\left(\hat{m}^* \geq n^{\gamma_1}\right) \leq c_D n^{-D}, \text{ i.e., } \mathbb{P}\left(\hat{m}^* \leq n^{\gamma_1}\right) \geq 1 - c_D n^{-D}.$$

Thus we have

$$\mathbb{P}\left(n^{\gamma_1} \geq \hat{m}^* \geq n^{\gamma}\right) \geq 1 - c_D n^{-D}.$$

4.3.2 Proof of Lemma 2

Let $\mathcal{F}_j = \left\{\frac{1}{2}\left|\theta_j - \theta_k\right| \leq \left|\hat{\theta}_j - \theta_k\right| \leq 2\left|\theta_j - \theta_k\right|, k \neq j\right\}$, $j \leq n^{\gamma_1}$. From the assumption (18) we have $|\theta_j - \theta_k| \geq M_0^{-1} n^{-(\alpha+1)\gamma_1}$ with $(\alpha+1)\gamma_1 < \frac{1}{2}$. Then equations (34) and (35) imply that for any $D > 0$ there exists a constant $C_D > 0$ such that for $j \leq n^{\gamma_1}$

$$\mathbb{P}\left(\mathcal{F}_j^c\right) \leq c_D n^{-D} \tag{36}$$

and consequently

$$\mathbb{P}\left(\cup_{j\leq n^{\gamma_1}, k\neq j}\left\{\frac{1}{2}\left|\theta_j - \theta_k\right| \leq \left|\hat{\theta}_j - \theta_k\right| \leq 2\left|\theta_j - \theta_k\right|\right\}\right) \geq 1 - c_D n^{-D}. \tag{37}$$

Note that

$$\hat{\phi}_j - \phi_j = \sum_k \phi_k \int\left(\hat{\phi}_j - \phi_j\right)\phi_k = \sum_{k:k\neq j}\phi_k \int \hat{\phi}_j\phi_k + \phi_j\int\left(\hat{\phi}_j\phi_j - 1\right).$$

The facts $\int \hat{K}(u,v)\hat{\phi}_j(u)\,du = \hat{\theta}_j\hat{\phi}_j(v)$ and $\int K(u,v)\phi_k(v)\,dv = \theta_k\phi_k(u)$ imply

$$\int \hat{\phi}_j\phi_k = \left(\hat{\theta}_j - \theta_k\right)^{-1}\int\int \hat{K}(u,v) - K(u,v)\hat{\phi}_j(u)\,\phi_k(v)\,dudv.$$

Now it follows from the elementary inequality $1 - x \leq \sqrt{1-x} \leq 1 - x/2$ for $0 \leq x \leq 1$ (we assume that $\int \hat{\phi}_j\phi_j \geq 0$ WLOG) that

$$1 - \sum_{k\neq j}\left[\int \hat{\phi}_j\phi_k\right]^2 \leq \int \hat{\phi}_j\phi_j = \sqrt{1 - \sum_{k\neq j}\left[\int \hat{\phi}_j\phi_k\right]^2} \leq 1 - \frac{1}{2}\sum_{k\neq j}\left[\int \hat{\phi}_j\phi_k\right]^2.$$

Then we have

$$\left\|\hat{\phi}_j - \phi_j\right\|^2 \leq 2\sum_{k:k\neq j}\left[\left(\hat{\theta}_j - \theta_k\right)^{-1}\int\int\left(\hat{K}(u,v) - K(u,v)\right)\hat{\phi}_j(u)\,\phi_k(v)\,dudv\right]^2$$

which on $\mathcal{F}_j$ is further bounded by

$$\begin{aligned}
& 8\sum_{k:k\neq j}\left[(\theta_j-\theta_k)^{-1}\int\int\left(\hat{K}(u,v)-K(u,v)\right)\hat{\phi}_j(u)\,\phi_k(v)\,dudv\right]^2 \\
\leq\ & 16\sum_{k:k\neq j}(\theta_j-\theta_k)^{-2}\left\{\begin{array}{c}\left[\int\int\left(\hat{K}(u,v)-K(u,v)\right)\left(\hat{\phi}_j(u)-\phi_j(u)\right)\phi_k(v)\,dudv\right]^2 \\ +\left[\int\int\left(\hat{K}(u,v)-K(u,v)\right)\phi_j(u)\,\phi_k(v)\,dudv\right]^2\end{array}\right\} \\
\leq\ & Cn^{2\gamma_1(\alpha+1)}\Delta^2\left\|\hat{\phi}_j-\phi_j\right\|^2+16\sum_{k:k\neq j}(\theta_j-\theta_k)^{-2}\left[\int\int\left(\hat{K}(u,v)-K(u,v)\right)\phi_j(u)\,\phi_k(v)\,dudv\right]^2.
\end{aligned}$$

This implies for each $D>0$

$$\mathbb{P}\left(\frac{1}{2}\left\|\hat{\phi}_j-\phi_j\right\|^2\leq 16\sum_{k:k\neq j}(\theta_j-\theta_k)^{-2}\left[\int\int\left(\hat{K}(u,v)-K(u,v)\right)\phi_j(u)\,\phi_k(v)\,dudv\right]^2\right)\geq 1-c_Dn^{-D}.$$

Let $\eta_{i,j}=\int X_i\phi_j$ and $\overline{\eta}_j=\frac{1}{n}\sum_i\eta_{i,j}$, then

$$X_i-\bar{X}=\sum_{j=1}^{\infty}\left(\eta_{i,j}-\overline{\eta}_j\right)\phi_j.$$

Assume without loss of generality that $\mathbb{E}X=0$ and for $k\neq j$ write

$$\int\int\left[\hat{K}(u,v)-K(u,v)\right]\phi_j(u)\,\phi_k(v)\,dudv=\frac{1}{n}\sum_{i=1}^{n}\left(\eta_{i,j}-\overline{\eta}_j\right)\left(\eta_{i,k}-\overline{\eta}_k\right)=\frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}\eta_{i,k}-\overline{\eta}_k\overline{\eta}_j$$

where $\frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}\eta_{i,k}$ is the dominating term. From the assumption (20) we have

$$\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\eta_{i,j}\eta_{i,k}\right)^2\leq n^{-1}\mathbb{E}\left(\eta_{1,j}\eta_{1,k}\right)^2\leq n^{-1}\left[\mathbb{E}\eta_{1,j}^4\eta_{1,k}^4\right]^{1/2}\leq C_1n^{-1}\theta_j\theta_k.$$

Note that the spacing condition in (18) implies $\theta_m-\theta_{2m}\asymp m^{-\alpha}$, so we have

$$\begin{aligned}
\mathbb{E}\left\|\hat{\phi}_j-\phi_j\right\|^2 &\leq C\sum_{k:k\neq j}(\theta_j-\theta_k)^{-2}n^{-1}\theta_j\theta_k \\
&\leq Cn^{-1}\theta_j\sum_{k:k\neq j}\left\{j^{2\alpha}\sum_{k:k\geq 2j}k^{-\alpha}+\sum_{k:k\leq j/2}k^{\alpha}+j^{2(\alpha+1)}\sum_{k:2j\geq k\geq j/2}\frac{k^{-\alpha}}{(1+|j-k|)^2}\right\} \\
&\leq C_1n^{-1}j^2 \qquad (38)
\end{aligned}$$

and the first part of lemma is proved.

For the second part of the lemma, equation (38) implies that it suffices to show that for $j \leq n^{\gamma_1}$ and all $\delta > 0$

$$\mathbb{P}\left(\cup_k \left\{ n^{1-\delta} k^\alpha j^\alpha \left[\int\int \left(\hat{K}(u,v) - K(u,v) \right) \phi_j(u) \phi_k(v) \, du dv \right]^2 \geq 1 \right\} \right) \leq c_D n^{-D}. \quad (39)$$

For a large constant $q > 0$, we have

$$\begin{aligned} & \mathbb{E} \sum_{k > n^q} (\theta_j - \theta_k)^{-2} \left[\int\int \left(\hat{K}(u,v) - K(u,v) \right) \phi_j(u) \phi_k(v) \, du dv \right]^2 \\ \leq \; & C \mathbb{E} \frac{\theta_j^{-2}}{n^2} \sum_{k > n^q} \left(\sum_{i=1}^n \eta_{i,j} \eta_{i,k} \right)^2 \leq C_1 \theta_j^{-1} n^{-1} \theta_k \leq C_q \theta_j^{-1} n^{-1} n^{-q\alpha}, \end{aligned}$$

which can be smaller than n^{-D} by setting q sufficiently large. It follows from the Markov inequality that

$$\mathbb{P}\left(\cup_{k > n^q} \left\{ n^{1-\delta} k^\alpha j^\alpha \left[\int\int \left(\hat{K}(u,v) - K(u,v) \right) \phi_j(u) \phi_k(v) \, du dv \right]^2 \geq 1 \right\} \right) \leq c_D n^{-D}.$$

We need now only to consider $k \leq n^q$. Let w be a positive integer. Then

$$\mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n \eta_{i,j} \eta_{i,k} \right)^{2w} \leq n^{-w} \mathbb{E}\left(\eta_{1,j} \eta_{1,k} \right)^{2w} \leq n^{-w} \left[\mathbb{E} \eta_{1,j}^{4w} \eta_{1,k}^{4w} \right]^{1/2} \leq C_1 n^{-w} \theta_j^w \theta_k^w$$

where the last inequality follows from (20). The Markov Inequality yields that for every integer $k > 0$

$$\mathbb{P}\left\{ n^{1-\delta} k^\alpha j^\alpha \left[\int\int \left(\hat{K}(u,v) - K(u,v) \right) \phi_j(v) \phi_k(v) \, du dv \right]^2 \geq 1 \right\} \leq C_2 n^{-w\delta}.$$

By choosing w sufficiently large, this implies

$$\mathbb{P}\left(\cup_{k \leq n^q} \left\{ n^{1-\delta} k^\alpha j^\alpha \left[\int\int \left(\hat{K}(u,v) - K(u,v) \right) \phi_j(u) \phi_k(v) \, du dv \right]^2 \geq 1 \right\} \right) \leq c_D n^{-D}.$$

The equation (39) is then proved, and so is the second part of the lemma.

4.3.3 Proof of Lemmas 3 and 4

Since $\mathbb{V}ar(\check{b}_j) \leq \mathbb{E}(\int b\hat{\phi}_j - \int b\phi_j)^2$, we will analyze $\int b\hat{\phi}_j - \int b\phi_j = \int b\left(\hat{\phi}_j - \phi_j\right)$ instead. By the Cauchy-Schwarz inequality we have

$$\mathbb{E}\left[\int b\left(\hat{\phi}_j - \phi_j\right) \right]^2 \leq CE \left\| \hat{\phi}_j - \phi_j \right\|^2 \leq C_1 j^2/n = o\left(\frac{j^\alpha}{n} \right). \quad (40)$$

We need to analyze $\tilde{d}_j = \hat{\theta}_j^{-\frac{1}{2}} \tilde{g}_j$. It follows from (12) that

$$\tilde{d}_j = \hat{\theta}_j^{-\frac{1}{2}} \tilde{g}_j = \hat{\theta}_j^{\frac{1}{2}} \check{b}_j + \hat{\theta}_j^{-\frac{1}{2}} \frac{1}{n} \hat{x}'_{\cdot,j} (\underline{\mathrm{Z}} - \bar{Z}).$$

Hence, $\mathbb{E}(\tilde{d}_j) = \mathbb{E}(\hat{\theta}_j^{\frac{1}{2}} \check{b}_j)$. Same as before, it follows from the fact $\mathbb{V}ar(Y) = \mathbb{E}(\mathbb{V}ar(Y|X)) + \mathbb{V}ar(\mathbb{E}(Y|X))$ for any two random variables X and Y that

$$\mathbb{V}ar(\tilde{d}_j) = \mathbb{V}ar(\hat{\theta}_j^{\frac{1}{2}} \check{b}_j) + \frac{\sigma^2}{n^2} \sum_{i=1}^{n} \mathbb{E}(\hat{\theta}_j^{-1} \hat{x}_{i,j}^2) = \mathbb{V}ar(\hat{\theta}_j^{\frac{1}{2}} \check{b}_j) + \frac{\sigma^2}{n}.$$

We need to bound $\mathbb{V}ar(\hat{\theta}_j^{\frac{1}{2}} \check{b}_j)$. Note that

$$\begin{aligned}
\mathbb{V}ar(\hat{\theta}_j^{\frac{1}{2}} \check{b}_j) &\leq \mathbb{E} \left(\hat{\theta}_j^{\frac{1}{2}} \check{b}_j - \theta_j^{1/2} b_j \right)^2 \\
&\leq 2\mathbb{E} \left(\hat{\theta}_j^{\frac{1}{2}} - \theta_j^{1/2} \right)^2 b_j^2 + 2\theta_j \mathbb{E} \left(\check{b}_j - b_j \right)^2 \\
&\leq 2\mathbb{E} \left(\hat{\theta}_j^{\frac{1}{2}} - \theta_j^{1/2} \right)^2 b_j^2 + Cn^{-1} j^{2-\alpha} \\
&\leq 2\mathbb{E} \left(\frac{\hat{\theta}_j - \theta_j}{\theta_j^{1/2}} \right)^2 b_j^2 + Cn^{-1} j^{2-\alpha} \\
&\leq Cn^{-1} j^{-2\beta+\alpha} + Cn^{-1} j^{2-\alpha} \leq C_1 n^{-1} j^{2-\alpha}. \qquad (41)
\end{aligned}$$

Here the third inequality follows from (40).

4.3.4 Proof of Lemma 5

Recall that

$$\tilde{d}_j = \hat{\theta}_j^{-1/2} \tilde{g}_j = \hat{\theta}_j^{1/2} \check{b}_j + \hat{\theta}_j^{-1/2} \frac{1}{n} \hat{x}'_{\cdot,j} (\underline{\mathrm{Z}} - \bar{Z}).$$

The second term is dominant. We consider this term first. Since

$$\frac{1}{n} \sum_{i=1}^{n} \hat{x}_{i,j} \hat{x}_{i,k} = \hat{\theta}_j \delta_{j,k},$$

we have

$$\sum_{j=m_1}^{m_2} \left[\hat{\theta}_j^{-1/2} \frac{1}{\sqrt{n}} \hat{x}'_{\cdot,j} \underline{\mathrm{Z}} \right]^2 \sim \frac{\sigma^2}{n} \chi^2_{m_2 - m_1 + 1}.$$

So for any $D > 0$ there exists a constant $C_D > 0$ such that

$$\mathbb{P} \left(\sum_{j=m_1}^{m_2} \hat{\theta}_j^{-1} \left[\frac{1}{n} \hat{x}'_{\cdot,j} (\underline{\mathrm{Z}} - \bar{Z}) \right]^2 > (1+\varepsilon) (m_2 - m_1) \frac{\sigma^2}{n} \right) \leq C_D n^{-D}. \qquad (42)$$

Now we turn to the first term. It is easy to see

$$\sum_{j=m_1}^{m_2} \theta_j b_j^2 \leq \varepsilon \frac{m_2 - m_1}{n},$$

and for any $D > 0$

$$\mathbb{P}\left(\left|\hat{\theta}_j - \theta_j\right| \geq \varepsilon\theta_j,\ j \leq n^{\gamma_1}\right) \leq C_D n^{-D}.$$

We need only to show that for any $D > 0$

$$\mathbb{P}\left(\sum_{j=m_1}^{m_2} \theta_j \left[\int b\left(\hat{\phi}_j - \phi_j\right)\right]^2 > \varepsilon\left(m_2 - m_1\right)\frac{\sigma^2}{n}\right) \leq C_D n^{-D}.$$

By the Cauchy-Schwarz inequality it suffices to show that for any $D > 0$

$$\mathbb{P}\left(\theta_j \int \left(\hat{\phi}_j - \phi_j\right)^2 > \varepsilon\frac{\sigma^2}{n}\right) \leq C_D n^{-D}. \tag{43}$$

This follows directly from Lemma 2.

4.4 Proof of Lemma 6

We write

$$\begin{aligned}
\sum_{j\in B_i} \epsilon_j^2 &= \sum_{j\in B_i}\left(\tilde{d}_j - d'_j\right)^2 = \sum_{j\in B_i}\left[\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j + \hat{\theta}_j^{-\frac{1}{2}}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{\mathrm{Z}} - \bar{Z})\right]^2 \\
&= \sum_{j\in B_i}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)^2 + 2\sum_{j\in B_i}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)\hat{\theta}_j^{-\frac{1}{2}}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{\mathrm{Z}} - \bar{Z}) + \sum_{j\in B_i}\left[\hat{\theta}_j^{-\frac{1}{2}}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{\mathrm{Z}} - \bar{Z})\right]^2 \\
&\leq \sum_{j\in B_i}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)^2 + 2\left\{\sum_{j\in B_i}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)^2 \sum_{j\in B_i}\left[\hat{\theta}_j^{-\frac{1}{2}}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{\mathrm{Z}} - \bar{Z})\right]^2\right\}^{1/2} \\
&\quad + \sum_{j\in B_i}\left[\hat{\theta}_j^{-\frac{1}{2}}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{\mathrm{Z}} - \bar{Z})\right]^2
\end{aligned}$$

We first show equation (28). From equation (42) it suffices to prove that, when $\lambda = 1+\varepsilon$ and $L_i \equiv |B_i| \geq n^\delta$ for some $\delta > 0$,

$$\mathbb{P}\left\{\sum_{j\in B_i}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)^2 > \frac{\varepsilon}{3}L_i\frac{\sigma^2}{n}\right\} \leq c_D n^{-D}$$

for any $D > 0$ where $C_D > 0$ is a constant. Note that, when $j \leq n^{\gamma_1}$, for any $D > 0$ there exists a constant $C_D > 0$ such that

$$\mathbb{P}\left(\left|\hat{\theta}_j - \theta_j\right| \geq \varepsilon^2\theta_j\right) \leq C_D n^{-D}$$

and

$$\mathbb{E}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)^2 = o\left(\frac{1}{n}\right) \text{ as } j \to \infty.$$

It then suffices to show that for all $D > 0$

$$\mathbb{P}\left(\sum_{j\in B_i}\theta_j\left[\int b\left(\hat{\phi}_j - \phi_j\right)\right]^2 > \varepsilon L_i\frac{\sigma^2}{n}\right) \le C_D n^{-D}.$$

This is true following similar arguments as in the proof of Lemma 5 with $L_i \ge n^\delta$ for some $\delta > 0$.

Equation (29) follows easily from the fact

$$\mathbb{E}\sum_{j\in B_i}\epsilon_j^2 = \mathbb{E}\sum_{j\in B_i}\left(\hat{\theta}_j^{\frac{1}{2}}\check{b}_j - d'_j\right)^2 + \mathbb{E}\sum_{j\in B_i}\left[\hat{\theta}_j^{-\frac{1}{2}}\frac{1}{n}\hat{x}'_{\cdot,j}(\underline{\mathrm{Z}} - \bar{Z})\right]^2$$

where the first term is bounded by $\frac{C}{n}L_i$ from equation (41) and the second term is exactly $\frac{\sigma^2}{n}L_i$.

References

[1] Bhatia, R., Davis, C. and McIntosh, A. (1983). Perturbation of spectral subspaces and solution of linear operator equations. *Linear Algebra Appl.* ***52/53***, 45-67.

[2] Brown, L. D., Cai, T. T., Zhang, R., Zhao, L. H. and Zhou, H. H. (2007). The Root-unroot algorithm for density estimation as implemented via wavelet block thresholding. Submitted.

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

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

[5] Cai, T., Low, M., and Zhao, L. (2000). Sharp adaptive estimation by a blockwise method. Unpublished manuscript.

[6] Cardot and Sarda, P. (2003). Linear regression models for functional data. Manuscript.

[7] 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.

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

[9] 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.

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

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

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

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

[14] Li, Y. and Hsing, T. (2007). On rates of convergence in functional linear regression. *J. Multivariate Analysis* ***98***, 1782-1804.

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

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

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