Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 54 (3) (2025), 1107 – 1127
DOI : 10.15672/hujms.1608956

Research Article

Smoothed least absolute deviation estimation in functional linear model

Yanfei He, Ping Yu*, Jianhong Shi, Wenhui Xuan

School of Mathematics and Computer Science, Shanxi Normal University, Tai Yuan 030031, China

Abstract

The functional linear model extends classical regression by modeling scalar responses as functions of stochastic processes. This paper proposes a novel convolution-type smoothed least absolute deviation estimator that addresses the non-smoothness and strict convexity challenges of conventional least absolute deviation estimation. By approximating both the predictor variable and slope function via functional principal component basis expansions, we develop a robust estimator with strong theoretical guarantees. Under mild regularity conditions, we establish the estimator's consistency aligning with the least absolute deviation estimator as the bandwidth vanishes and derive the convergence rate for the prediction error. Simulation studies demonstrate that the proposed smoothed least absolute deviation estimator outperforms conventional estimation methods–including ordinary least squares, standard least absolute deviation, spline-based regression, penalized spline smoothing, and Bayesian estimation, particularly in scenarios involving heavy-tailed error distributions, outlier contamination, and heteroscedasticity. Applications to the Berkeley Growth Study and the Capital Bike Share dataset further validate its practical utility.

Mathematics Subject Classification (2020). 62G05, 62G07

Keywords. Convolution-type smoothed least absolute deviation, functional principal component analysis, least absolute deviation, robust estimation

1. Introduction

Advances in data collection and storage have tremendously increased the presence of functional data, whose graphical representations are curves [1], shapes [2], or images [3], etc. For a comprehensive treatment on the subject of functional data analysis, we recommend the monographs [4–9]. Functional regression analysis is one of the most useful techniques in functional data analysis. The basic idea behind functional regression analysis is to regard longitudinally observed predictors and/or responses as smooth functional data, and then elucidate the relationship between the responses and predictors and predict the newly observed data through the estimated model. As a fundamental tool for assessing the relationship between two random variables of functional or scalar form, FLM has been extensively investigated in the literature [10–17], among others.

*Corresponding Author.
Email addresses: 15535188987@163.com (Y. He), ypxye@sina.com (P. Yu), shijh70@163.com (J. Shi), xuanwh99@163.com (W. Xuan)
Received: 28.12.2024; Accepted: 08.05.2025

The early functional linear model (FLM) mainly relied on the OLS method. Although OLS performs optimally under Gaussian assumptions, real-life functional data frequently deviate from these conditions. Such deviations manifest as (i) heavy-tailed errors in growth curve analysis [18] and ophthalmological data [3] due to biological variability; (ii) demand anomalies and extreme traffic flows in urban mobility systems [19, 20]; and (iii) extreme events in environmental monitoring [21] and econometric forecasting [8].

When conditional median is of primary interest or the data contain substantial outliers, both least absolute deviation (LAD) [22] and quantile regression (QR) [23] offer more robust alternatives than mean regression. However, their non-differentiable loss functions present significant computational challenges in functional contexts, particularly when requiring: (i) dimension reduction through basis expansions; (ii) enforcement of smoothness constraints; and (iii) large-scale optimization. To address these limitations, we developed a smoothed least absolute deviation (SLAD) estimator for FLM that incorporates recent advances in smoothing methodologies [24–26]. Our approach provides three key advantages: (i) enhanced robustness against data contamination in growth curve analysis [18]; (ii) improved computational efficiency compared to conventional LAD; and (iii) greater flexibility in handling irregular sampling and high-dimensional predictors.

This work bridges a critical gap between LAD's robustness and functional data's dimensionality requirements by addressing three fundamental challenges. Theoretically, it reconciles robust statistics with infinite-dimensional functional spaces through: (i) constructing a smoothed LAD objective function approximation; (ii) employing functional principal component analysis for dimension reduction; and (iii) establishing asymptotic properties that confirm matching convergence rates with standard LAD estimation as bandwidth vanishes while deriving the convergence rate for the prediction error. Methodologically, the framework resolves the robustness-smoothness trade-off through three integrated components: a quadratically differentiable loss function, robust FPCA, and an efficient MM algorithm. Empirical validation through benchmark datasets demonstrates both methodological efficacy and practical value. The Berkeley Growth study reveals systematic variations in height prediction accuracy by biological sex and developmental phase, while the Capital Bike Share analysis identifies an inverted U-shaped temperature-usage relationship. In particular, SLAD outperforms conventional approaches (OLS, LAD, spline methods [27], penalized spline [28] and Bayesian alternatives [29]) by simultaneously achieving: (i) superior robustness against heavy-tailed distributions, outliers, and heteroscedastic errors; (ii) maintained predictive accuracy, and (iii) computational feasibility. Collectively, this work represents the first unified solution that bridges the long-standing divide between LAD's robustness and functional data's dimensionality requirements while delivering three key advances: heavy-tailed error robustness, smoothness preservation, and practical computational efficiency.

The paper is organized as follows: Section 2 develops an estimation method for the SLAD and derives the asymptotic properties of the proposed estimator. An algorithm based on majorize-minimization is detailed in Section 3. Section 4 performs numerical simulations with a finite sample. Section 5 applies the proposed method to the Berkeley Growth data and the Capital Bike Share datasets. Section 6 concludes the paper with some discussion.

2. Methodology and main conclusions

In this section, we first give a brief introduction to the SLAD technique based on the classical linear regression model, and then extend this method to the FLM.

2.1. Methodology

Let $Y \in \mathbb{R}$ be a univariate response variable, $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^\top \in \mathbb{R}^p$ be the p-dimensional covariate vector. Then, let $\{(Y_i, \boldsymbol{X}_i)\}_{i=1}^n$ be n independent and identically distributed (i.i.d.) random observations that are sampled from the following linear regression model

$$Y = \boldsymbol{X}^\top \boldsymbol{\beta}^* + \varepsilon, \tag{2.1}$$

where $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*, \ldots, \beta_p^*)^\top \in \mathbb{R}^p$ is the true parameter and ε is the random noise. Assume that ε is independent of $\boldsymbol{X}$, and has unknown cumulative distribution function $F_\varepsilon(\cdot)$ with f_ε as its density function. The LAD estimator $\hat{\boldsymbol{\beta}}$ is obtained by minimizing the random criterion

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \left| Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta} \right|. \tag{2.2}$$

For the above LAD estimator $\hat{\boldsymbol{\beta}}$, Pollard [22] provided a direct proof of asymptotic normality. However, the loss function lacks differentiability, which increases the computational burden and reduces the efficiency of statistical inference. To solve this problem, we introduce the SLAD method. After taking the expectation of the objective function in Equation (2.2) with respect to the true distribution of ε, we obtain the population-level objective function $\int_{-\infty}^{\infty} |t| \mathrm{d}F_\varepsilon(\cdot)$. Similar to [24] and [30], we will use the smooth estimator $\tilde{F}_\varepsilon(\cdot)$ in place of $F_\varepsilon(\cdot)$, then $\int_{-\infty}^{\infty} |t| \mathrm{d}\tilde{F}_\varepsilon(\cdot)$ can be the new objective function, which is smooth. Specifically, we consider the following kernel density estimator

$$\tilde{F}_\varepsilon(t) = \int_{-\infty}^{t} \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K\left(\frac{\nu - \varepsilon_i}{h} \right) \mathrm{d}\nu$$

with the kernel function $K : \mathbb{R} \to [0, \infty)$ and bandwidth $h > 0$. The new objective function is

$$\int_{-\infty}^{\infty} |t| \mathrm{d}\tilde{F}_\varepsilon(t) = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{\varepsilon_i - t}{h} \right) |t| \mathrm{d}t := \frac{1}{n} \sum_{i=1}^n L_h(\varepsilon_i),$$

where $L_h(u) = \int_{-\infty}^{\infty} |\nu| \frac{1}{h} K(\frac{u-\nu}{h}) \mathrm{d}\nu = \int_{-\infty}^{\infty} |u - \nu| \frac{1}{h} K(\frac{\nu}{h}) \mathrm{d}\nu$.

For the model in Equation (2.2), the SLAD estimator $\hat{\boldsymbol{\beta}}_h$ is defined as a solution to the following optimization problem

$$\arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n L_h(Y_i - \mathbf{X}_i^\top \boldsymbol{\beta}). \tag{2.3}$$

Next, we apply this method to the FLM, which is defined as

$$Y = \alpha_0 + \int_{\mathscr{T}} X(t) \beta(t) dt + \epsilon, \tag{2.4}$$

where Y is a real-valued scalar response, α_0 is the intercept term, $\beta(t) \in L^2(\mathscr{T})$ is the unknown slope function associated with functional predictor $X(t)$, ϵ is a random error independent of $X(t)$. Here, the Hilbert space $L^2(\mathscr{T})$ is the set of all square integrable functions on $\mathscr{T}$, endowed with inner product $\langle x, y \rangle = \int_{\mathscr{T}} x(t) y(t) dt$ and norm $\|x\| = \langle x, x \rangle^{1/2}$. Note that once an estimator $\hat{\beta}(t)$ of slope $\beta(t)$ is available, it is straightforward to estimate the intercept α_0, for example, as the average values of $Y - \int_{\mathscr{T}} X(t) \hat{\beta}(t) dt$. Therefore, much interest in the literature is focused on estimating $\beta(t)$. For simplicity of notation, we assume $\alpha_0 = 0$. Without loss of generality, we further assume that $\mathscr{T} = [0, 1]$, and $\{X(t) : t \in \mathscr{T}\}$ has zero mean throughout the study.

Let $(X_i(\cdot), Y_i), i = 1, 2, \ldots, n$ be the realizations of $(X(\cdot), Y)$ in model (2.4). The covariance and empirical covariance functions of $X(\cdot)$ can be defined as

$$C_X(s,t) = \mathrm{Cov}(X(s), X(t)), \quad \hat{C}_X(s,t) = \frac{1}{n}\sum_{i=1}^{n} X_i(s)X_i(t).$$

By the Mercer's Theorem, the spectral expansions of $C_X(s,t)$ and $\hat{C}_X(s,t)$ can be written as

$$C_X(s,t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s)\phi_j(t), \quad \hat{C}_X(s,t) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(s)\hat{\phi}_j(t),$$

where $\lambda_1 > \lambda_2 > \cdots > 0$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{n+1} = \cdots = 0$ are ordered nonnegative eigenvalues, $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ and $\{\hat{\phi}_j(\cdot)\}_{j=1}^{\infty}$ is continuous eigenfunctions of the covariance operator. With a slight abuse of notation, we use C_X to denote both the covariance operator and covariance function of $X(\cdot)$. We assume that the covariance operator C_X defined by $C_X f(s) = \int_0^1 C_X(s,t)f(t)dt$ is strictly positive. Obviously, $\{\phi_j(\cdot)\}_{j=1}^{\infty}$ are orthonormal basis functions on $L^2[0,1]$, and $\{\hat{\lambda}_j, \hat{\phi}_j(\cdot)\}$ can be regarded as estimators of $\{\lambda_j, \phi_j(\cdot)\}$. By the Karhunen-Loève representation, $\mathcal{L}^2$-valued functions $X(t)$ and $\beta(t)$ can be expanded to

$$X(t) = \sum_{j=1}^{\infty} \xi_j \phi_j(t), \quad \beta(t) = \sum_{j=1}^{\infty} \gamma_j \phi_j(t), \tag{2.5}$$

where $\xi_j = \langle X(\cdot), \phi_j(\cdot)\rangle$ is the jth score of $X(\cdot)$, and $\gamma_j = \langle \beta(\cdot), \phi_j(\cdot)\rangle$ is the jth fourier coefficient of $\beta(t)$, and the ξ_j are uncorrelated random variables with mean 0 and variance $\mathrm{E}\xi_j^2 = \lambda_j$. Analogously, we define $C_{YX} = \mathrm{Cov}(Y, X(\cdot))$, and empirical counterpart can be defined as $\hat{C}_{YX} = \frac{1}{n}\sum_{i=1}^{n} Y_i X_i$.

Given the orthogonality of $\{\phi_j(\cdot)\}_{j=1}^{m}$ and in Equation (2.5), the model in Equation (2.4) can be rewritten as:

$$Y_i = \sum_{j=1}^{m} \xi_{ij}\gamma_j + \tilde{\epsilon}_i, \qquad i = 1, 2, \ldots, n, \tag{2.6}$$

where $\xi_{ij} = \langle X_i(\cdot), \phi_j(\cdot)\rangle$, $\tilde{\epsilon}_i = \sum_{j=m+1}^{\infty} \xi_{ij}\gamma_j + \epsilon_i$, m is the "cutoff" level such that $1 \leq m \leq n-1$ and $m \to \infty$ as $n \to \infty$. Let $\boldsymbol{U}_i = (\xi_{i1}, \xi_{i2}, \ldots, \xi_{im})^\top$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_m)^\top$. The SLAD estimators $\hat{\beta}_h(t) = \sum_{j=1}^{m} \hat{\gamma}_{hj}\hat{\phi}_j(t)$ can be obtained by minimizing the loss function with respect to $\boldsymbol{\gamma}$ as follows:

$$\hat{Q}_h(\gamma) = \frac{1}{n}\sum_{i=1}^{n} L_h(Y_i - \hat{\boldsymbol{U}}_i^\top \boldsymbol{\gamma}), \tag{2.7}$$

where $\hat{\boldsymbol{U}}_i = (\hat{\xi}_{i1}, \hat{\xi}_{i2}, \ldots, \hat{\xi}_{im})^\top$ with $\hat{\xi}_{ij} = \langle X_i(t), \hat{\phi}_j(t)\rangle$.

Implementation of the proposed estimation methods requires the selection of the tuning parameter m. In this study, m is selected as the minimum value that reaches a certain proportion (denoted by ς) of the cumulative percentage of total variance (CPV) by the first m leading components as follows:

$$m = \arg\min_J \left\{ \sum_{j=1}^{J} \hat{\lambda}_j \Big/ \sum_{j=1}^{M} \hat{\lambda}_j \geqslant \varsigma \right\},$$

where M is the largest number of functional principal components, such that $\hat{\lambda}_j > 0$, and $\varsigma = 95\%$ is used in our numerical studies.

2.2. Main conclusions

In this section, we will study the large sample properties of the SLAD estimator defined in the previous section. In order to derive the asymptotic properties, we need the following conditions:

Condition 2.1 The errors $\{\epsilon_i\}_{i=1}^n$ are i.i.d. with density function $f_\epsilon(\cdot)$ about the Lebesgue measure on $\mathbb{R}$. We assume that the second-order derivative of $f_\epsilon(\cdot)$ exists on the real line and $\sup_{x\in\mathbb{R}}|f_\epsilon''(\cdot)| < \infty$.

Condition 2.2 Turning parameter m satisfied $m \sim n^{\frac{1}{a+2b}}$.

Condition 2.3 Random process $X(\cdot)$ and the score ξ_j satisfy conditions: $\mathrm{E}\|X(\cdot)\|^4 < \infty$, $\mathrm{E}[\xi_j^4]$
$\leq c\lambda_j^2$, $j \geq 1$.

Condition 2.4 For the eigenvalues λ_j of the linear operator C_X and and score coefficients γ_j, the following conditions hold:
(a) There exist constants c and $a > 1$, such that $c^{-1}j^{-a} \leq \lambda_j \leq cj^{-a}$, $\lambda_j - \lambda_{j-1} \geq cj^{-a-1}$, $j \geq 1$;
(b) There exist constants c and $b > 1 + \frac{a}{2}$, such that $|\gamma_j| \leq cj^{-b}$, $j \geq 1$.

Condition 2.5 $K(\cdot)$ that satisfies the following properties: (i) $K(-t) = K(t)$, $\forall t \in \mathbb{R}$; (ii) $\exists\delta_0$
> 0 *s.t.* $\kappa_l := inf_{t\in[-\delta_0,\delta_0]}K(t) > 0$; (iii) $\int_{-\infty}^{\infty} K(t)dt = 1$; (iv) $\kappa_u := sup_{t\in\mathbb{R}}K(t) < \infty$; (v) $\kappa_j := \int_{-\infty}^{\infty}|t|^j K(t)dt < \infty$, $j = 1, 2$; (vi) $\exists$ constant $\alpha_0 \in [0, 1]$, $L_0 > 0$ *s.t.* $|K(x) - K(y)| \leq L_0|x - y|^{\alpha_0}$ for any $x, y \in \mathbb{R}$.

Remark 2.1. Condition 2.1 is commonly required for LAD models. Conditions 2.2-2.4 are standard assumptions used in the classical functional linear regression (see, e.g., [11, 12]). In particular, Condition 2.2 gives the order of the truncated parameter m. Condition 2.3 is needed for the consistency of $\hat{C}_X$, and the second part of Condition 2.3 is satisfied if $X(\cdot)$ is a Gaussian process. Condition 2.4 (a) is required to identify the slope function $\beta(t)$ by preventing the spacings among the eigenvalues from being too small, while condition 2.4 (b) is used to make the slope function $\beta(t)$ sufficiently smooth. There can be a lot of choices for the kernel function $K(\cdot)$ satisfying Condition 2.5. For example, the Gaussian kernel with $K(u) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{u^2}{2}}$, the triangular kernel with $K(u) = (1 - |u|)\, I(-1 \leq u \leq 1)$ and the Epanechnikov kernel with $K(u) = \frac{3}{4}\,(1 - u^2)\, I(-1 \leq u \leq 1)$ all satisfy the conditions, where $I(\cdot)$ is the indicator function.

To facilitate the proof, we first give some additional notation. Let $\beta_0(t)$ be the true function of $\beta(t)$, and let $\boldsymbol{\gamma}_0 = (\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0m})^\top$ be the true coefficient coefficient vector of the score. Notation $\|\cdot\|$ denotes the $\mathscr{L}^2$ norm for a function or the Euclidean norm for a vector. In what follows, c denotes a generic positive constant that may take different values. Moreover, $a_n \sim b_n$ means that $|\frac{a_n}{b_n}|$ is bounded away from zero and infinity as $n \to \infty$. Let $\boldsymbol{\gamma}_h^* = \arg\min_{\gamma\in\mathbb{R}^m} \mathrm{E}L_h(Y - \hat{\boldsymbol{U}}^\top\boldsymbol{\gamma})$.

In our proofs, we will frequently need the following expressions for the loss function $L_h(\cdot)$ and its derivatives. Recall $L_h(u) = \int_{-\infty}^{\infty}|u - \nu|\frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu, u \in \mathbb{R}$. A direct calculation gives

$$L_h(u) = u\int_{-u}^{u}\frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu - 2\int_{-\infty}^{u}\frac{\nu}{h}K(\frac{\nu}{h})\mathrm{d}\nu,$$

$$L_h'(u) = 2\int_{-\infty}^{u}\frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu - 1 = 2\int_{0}^{u}\frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu, \tag{2.8}$$

$$L_h''(u) = \frac{2}{h}K(\frac{u}{h}), \forall u \in \mathbb{R}. \tag{2.9}$$

Lemma 2.2. *For any $t_1, t_2, t \in \mathbb{R}$, we have (i) $L_h'(-t) = -L_h'(t)$, (ii) $|L_h(t_1) - L_h(t_2)| \le |t_1 - t_2|$, $|L_h'(t_1) - L_h'(t_2)| \le \frac{2}{h}\kappa_u |t_1 - t_2|$ and (iii) $|L_h''(t_1) - L_h''(t_2)| \le \frac{2L_0}{h^{1+\alpha_0}}|t_1 - t_2|^{\alpha_0}$.*

Proof. (i) Note that $L_h'(t) = 2\int_{-\infty}^{t} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu - 1$ and $\int_{-\infty}^{\infty} K(t)\mathrm{d}t = 1$, so we have $L_h'(-t) = 2\int_{-\infty}^{-t} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu - 1 = 2(1 - \int_{-t}^{\infty} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu) - 1 = 1 - 2\int_{-t}^{\infty} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu = 1 - 2\int_{-\infty}^{t} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu = -L_h'(t)$, where the last equality is due to a change of variable and $K(-t) = K(t)$, $t \in \mathbb{R}$. Thus the first statement is proved.

(ii) By the property of the kernel function, we have $\int_{-\infty}^{\infty} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu = 1$. Since $L_h'(t) = 2\int_{-\infty}^{t} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu - 1$ and $K(t) \ge 0$ for all t, we have $-1 \le L_h'(t) \le 2\int_{-\infty}^{\infty} \frac{1}{h}K(\frac{\nu}{h})\mathrm{d}\nu - 1 = 1$. Then the $|L_h'(t)| \le 1$, $t \in \mathbb{R}$. The second statement follows from the mean value theorem.

(iii) Similarly, by the property of of kernel function, we have $0 \le L_h''(t) = \frac{2}{h}K(\frac{t}{h}) \le \frac{2}{h}\kappa_u$, which means $|L_h''(t)| \le \frac{2}{h}\kappa_u$. Thus, by the choice of kernel, we have $|L_h''(t_1) - L_h''(t_2)| = \frac{2}{h}|K(\frac{t_1}{h}) - K(\frac{t_2}{h})| \le \frac{2L_0}{h^{1+\alpha_0}}|t_1 - t_2|^{\alpha_0}$. Then the proof is finished. □

Let $\epsilon_i(\boldsymbol{\gamma}) = Y_i - \hat{\boldsymbol{U}}_i^\top \boldsymbol{\gamma}$. By Lemma 2.2, the gradient and the Hessian matrix of $\hat{Q}_h(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ are, respectively,

$$\nabla \hat{Q}_h(\boldsymbol{\gamma}) = -\frac{1}{n}\sum_{i=1}^{n} K_h(\epsilon_i(\boldsymbol{\gamma}))\hat{\boldsymbol{U}}_i, \quad \nabla^2 \hat{Q}_h(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h}K\left(\frac{\epsilon_i(\boldsymbol{\gamma})}{h}\right)\hat{\boldsymbol{U}}_i\hat{\boldsymbol{U}}_i^\top, \tag{2.10}$$

where $K_h(u) := 2\int_{-\infty}^{u} \frac{1}{h}K(\frac{t}{h})dt - 1$.

Theorem 2.3. *Under Conditions 2.1–2.5, as $n \to \infty$, (1) for any $h > 0$, we have $\|\boldsymbol{\gamma}_h^* - \boldsymbol{\gamma}_0\| = O(h^2)$, (2) further, when $h \sim n^{-\frac{a+2b-1}{4(a+2b)}}$, we have $\|\hat{\beta}_h(\cdot) - \beta_0(\cdot)\| = O_p\left(n^{-\frac{2b-1}{a+2b}}\right)$.*

Proof. Let $\delta_n = \sqrt{\frac{m}{n}}$, $\boldsymbol{V}_n = \delta_n^{-1}(\hat{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h^*)^\top$. Let $R_i = \int_0^1 X_i(t)\beta_0(t)dt - \hat{\boldsymbol{U}}_i^\top \boldsymbol{\gamma}_h^*$, $\mathcal{F}_n = \{\boldsymbol{V}_n : \|\boldsymbol{V}_n\| = L\}$, where L is a large enough constant, $\mathcal{N}_n = \{(X_i(\cdot), Y_i)\}_{i=1}^n$.

Firstly, similar to Lemma A.1 in [31], we have $\|\boldsymbol{\gamma}_h^* - \boldsymbol{\gamma}_0\| = O(h^2)$. Next, we show, for any given $\eta > 0$, there exists a sufficient large constant $L = L_\eta$ such that

$$P\left\{\inf_{\boldsymbol{V}_n \in \mathcal{F}_n} \hat{Q}_h(\boldsymbol{\gamma}_h^* + \delta_n \boldsymbol{V}_n) > \hat{Q}_h(\boldsymbol{\gamma}_h^*)\right\} \ge 1 - \eta. \tag{2.11}$$

This implies that with the probability of at least $1 - \eta$ there exists a local minimizer $\hat{\boldsymbol{\gamma}}_h$ in the ball $\left\{\boldsymbol{V}_n : \|\boldsymbol{V}_n\| \le L\right\}$, such that $\|\hat{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h^*\| = O_p(\delta_n)$. By invoking $\|\hat{\phi}_j - \phi_j\|^2 = O_p(n^{-1}j^2)$ (see, e.g., [32]), we have

$$\begin{aligned} |R_i|^2 &= \left|\int_0^1 X_i(t)\beta(t)dt - \hat{\boldsymbol{U}}_i^\top \gamma_h^*\right|^2 \le 2\left|\sum_{j=1}^{m}\langle X_i, \hat{\phi}_j - \phi_j\rangle\gamma_{hj}^*\right|^2 + 2\left|\sum_{j=m+1}^{\infty}\langle X_i, \phi_j\rangle\gamma_{hj}^*\right|^2 \\ &\triangleq 2A_1 + 2A_2. \end{aligned} \tag{2.12}$$

For A_1, by Conditions 2.2–2.4, and the Hölder inequality, we obtain

$$\begin{aligned} A_1 &= \left|\sum_{j=1}^{m}\langle X_i, \hat{\phi}_j - \phi_j\rangle\gamma_{hj}^*\right|^2 \le cm\sum_{j=1}^{m}\|\hat{\phi}_j - \phi_j\|^2\gamma_{hj}^{*2} \le cm\sum_{j=1}^{m} O_p(n^{-1}j^{2-2b}) = O_p(n^{\frac{a+4b-4}{a+2b}}) \\ &= o_p(\delta_n^2). \end{aligned}$$

As for A_2, given that $E\left\{\sum_{j=m+1}^{\infty}\langle X_i, \phi_j\rangle\gamma_{hj}^*\right\} = 0$,

$$\mathrm{Var}\left\{\sum_{j=m+1}^{\infty}\langle X_i, \phi_j\rangle\gamma_{hj}^*\right\} = \sum_{j=m+1}^{\infty}\lambda_j\gamma_{hj}^{*2} \le c\sum_{j=m+1}^{\infty} j^{-(a+2b)} = O(n^{-\frac{a+2b-1}{a+2b}}),$$

we have $A_2 = O_p(n^{-\frac{a+2b-1}{a+2b}}) = o_p(\delta_n^2)$. Taking these together, we obtain

$$|R_i|^2 = O_p(n^{-\frac{a+2b-1}{a+2b}}) = o_p(\delta_n^2). \tag{2.13}$$

Let $P_n(\boldsymbol{V}_n) = \hat{Q}_h(\boldsymbol{\gamma}_h^* + \delta_n \boldsymbol{V}_n) - \hat{Q}_h(\boldsymbol{\gamma}_h^*)$. By using Taylor series expansion,

$$\begin{aligned} P_n(\boldsymbol{V}_n) =& \hat{Q}_h(\boldsymbol{\gamma}_h^* + \delta_n \boldsymbol{V}_n) - \hat{Q}_h(\boldsymbol{\gamma}_h^*) \\ =& \delta_n \nabla \hat{Q}_h^\top(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n + \frac{1}{2}\delta_n^2 \boldsymbol{V}_n^\top \nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n\{1+o_p(1)\} \\ \geq& \frac{1}{2}\delta_n^2 \boldsymbol{V}_n^\top \nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n - \|\delta_n \nabla \hat{Q}_h^\top(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n\|, \end{aligned} \tag{2.14}$$

where the last inequality holds because the Hessian matrix $\nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)$ is positive definite. Next we consider the terms $\delta_n^2 \boldsymbol{V}_n^\top \nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n$ and $\delta_n \nabla \hat{Q}_h^\top(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n$, respectively. Note that $\{\epsilon_i\}_{i=1}^n$ be the independent random variables, by the Taylor series expansion, we have

$$\begin{aligned} \|\nabla \hat{Q}_h(\boldsymbol{\gamma}_h^*)\| =& \left\| \frac{1}{n}\sum_{i=1}^n K_h\left\{\epsilon_i + R_i\right\} \hat{\boldsymbol{U}}_i^\top \right\| \\ =& \left\| \frac{1}{n}\sum_{i=1}^n \left\{ K_h(\epsilon_i) + K_h'(\epsilon_i) R_i + K_h''(\epsilon_i) R_i^2 + o_p(1) \right\} \hat{\boldsymbol{U}}_i^\top \right\| \\ \leq& \sqrt{3} \left\| \frac{1}{n}\sum_{i=1}^n K_h(\epsilon_i) \hat{\boldsymbol{U}}_i^\top \right\| + \sqrt{3} \left\| \frac{1}{n}\sum_{i=1}^n K_h'(\epsilon_i) R_i \hat{\boldsymbol{U}}_i^\top \right\| \\ &+ \sqrt{3} \left\| \frac{1}{n}\sum_{i=1}^n K_h''(\epsilon_i) R_i^2 \hat{\boldsymbol{U}}_i^\top + o_p(1) \right\|. \end{aligned}$$

According to the definition and conditions on kernel function, and Condition 2.1, we have $\mathrm{E}[K(\frac{\epsilon}{h})] = \int_{-\infty}^{\infty} K(\frac{\nu}{h}) f_\epsilon(\nu) \mathrm{d}\nu$. Let $\nu = th$, applying Taylor's Theorem, $\mathrm{E}[K(\frac{\epsilon}{h})] = h\int_{-\infty}^{\infty} K(t) f(th) dt = h\int_{-\infty}^{\infty} K(t)[f(0) + o(1)]dt$. Therefore, $\mathrm{E}|K_h(\epsilon)| = O(1)$. Similarly, $K_h'(\epsilon_i) = \frac{1}{2} f(0) + O_p(h^2)$. Combining $\mathrm{E}|K_h(\epsilon)| = O(1)$ with Conditions 2.1–2.3, one has

$$\mathrm{E}\left\| \frac{1}{n}\sum_{i=1}^n K_h(\epsilon_i) \hat{\boldsymbol{U}}_i^\top \right\|^2 = \frac{1}{n}\mathrm{E}\left\{ [K_h(\epsilon_1)]^2 \right\} \mathrm{E}\left\{ \hat{\boldsymbol{U}}_1 \hat{\boldsymbol{U}}_1^\top \right\} = \frac{1}{n}\{O(1)\} \sum_{j=1}^m \mathrm{E}\left\{ \hat{\boldsymbol{U}}_{1j}^2 \right\} = O\left(\frac{m}{n}\right).$$

Further, by (2.13) and Conditions 2.1–2.3, we have $\|\frac{1}{n}\sum_{i=1}^n K_h'(\epsilon_i) R_i \hat{\boldsymbol{U}}_i^\top\| = O_p\left(\frac{m}{n}\right)$. Similarly, we can obtain $\|\frac{1}{n}\sum_{i=1}^n K_h''(\epsilon_i) O_p(R_i^2) \hat{\boldsymbol{U}}_i^\top\| = o_p\left(\frac{m}{n}\right) = o_p(1)$. Hence, $\|\nabla \hat{Q}_h(\boldsymbol{\gamma}_h^*)\| = O_p\left(\sqrt{\frac{m}{n}}\right)$. In summary, we get

$$\|\delta_n \nabla \hat{Q}_h^\top(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n\| \leq \delta_n \|\nabla \hat{Q}_h(\boldsymbol{\gamma}_h^*)\| \|\boldsymbol{V}_n\| \leq O_p\left(\delta_n^2\right) \|\boldsymbol{V}_n\|.$$

For $\boldsymbol{V}_n^\top \nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n$, it is obvious that the Hessian matrix $\nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)$ is positive definite. By Condition 2.3, one has

$$\begin{aligned} \mathrm{E}\left\{ \boldsymbol{V}_n^\top \nabla^2 \hat{Q}_h(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n \right\} =& \boldsymbol{V}_n^\top \frac{1}{nh}\sum_{i=1}^n \mathrm{E}\left\{ \left(K\left(\frac{\epsilon_i}{h}\right) + K'\left(\frac{\epsilon_i}{h}\right) R_i \right) \hat{\boldsymbol{U}}_i \hat{\boldsymbol{U}}_i^\top \right\} \boldsymbol{V}_n \\ =& \boldsymbol{V}_n^\top \frac{1}{h}\mathrm{E}\left\{ K\left(\frac{\epsilon_1}{h}\right) \hat{\boldsymbol{U}}_1 \hat{\boldsymbol{U}}_1^\top \right\} \boldsymbol{V}_n + \boldsymbol{V}_n^\top \frac{1}{h}\mathrm{E}\left\{ K'\left(\frac{\epsilon_1}{h}\right) R_i \hat{\boldsymbol{U}}_1 \hat{\boldsymbol{U}}_1^\top \right\} \boldsymbol{V}_n. \end{aligned}$$

For $\boldsymbol{V}_n^\top \frac{1}{h}\mathrm{E}\left\{K\left(\frac{\epsilon_1}{h}\right)\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right\}\boldsymbol{V}_n$, since ϵ_i is independent of $\hat{\boldsymbol{U}}_i$, it follows that

$$\begin{aligned}
&\boldsymbol{V}_n^\top \frac{1}{h}\mathrm{E}\left\{K\left(\frac{\epsilon_1}{h}\right)\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right\}\boldsymbol{V}_n\\
\geq&\mathrm{E}\left\{\frac{1}{h}K\left(\frac{\epsilon_1}{h}\right)\right\}\lambda_{\min}\left\{\mathrm{E}\left(\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right)\right\}\|\boldsymbol{V}_n\|^2\\
=&\frac{1}{h}\int K\left(\frac{\epsilon_1}{h}\right)f(\epsilon_1)d(\epsilon_1)\lambda_{\min}\left\{\mathrm{E}\left(\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right)\right\}\|\boldsymbol{V}_n\|^2\\
\geq&\int K(t)\left\{f(0)+thf'(0)+\frac{(th)^2}{2}f''(0)\right\}dt\lambda_{\min}\left\{\mathrm{E}\left(\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right)\right\}\|\boldsymbol{V}_n\|^2\\
=&f(0)\lambda_{\min}\left\{\mathrm{E}\left(\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right)\right\}\|\boldsymbol{V}_n\|^2+o(h).
\end{aligned}$$

Similarly, one has

$$\begin{aligned}
\boldsymbol{V}_n^\top \frac{1}{h}\mathrm{E}\left\{K'\left(\frac{\epsilon_i}{h}\right)R_i\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right\}\boldsymbol{V}_n \geq c\lambda_{\min}\left\{\mathrm{E}\left(\hat{\boldsymbol{U}}_1\hat{\boldsymbol{U}}_1^\top\right)\right\}\mathrm{E}\{R_i\}\|\boldsymbol{V}_n\|^2 = O\left(\sqrt{\frac{m}{n}}\right)\|\boldsymbol{V}_n\|_2^2\\
=o(1)\|\boldsymbol{V}_n\|^2.
\end{aligned}$$

Thus, $\delta_n^2\boldsymbol{V}_n^\top\nabla^2\hat{Q}_h(\boldsymbol{\gamma}_h^*)\boldsymbol{V}_n \geq O_p\left(\delta_n^2\right)\|\boldsymbol{V}_n\|^2 + o_p\left(\delta_n^2\right)\|\boldsymbol{V}_n\|^2 = O_p\left(\delta_n^2\right)\|\boldsymbol{V}_n\|^2$. To sum up, $P_n(\boldsymbol{V}_n)$ is controlled by the term $O_p\left(\delta_n^2\right)\|\boldsymbol{V}_n\|^2$. Hence, the equation (2.11) holds, there exists local minimizer $\hat{\boldsymbol{\gamma}}_h$ such that

$$\|\hat{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h^*\| = O_p(\delta_n). \tag{2.15}$$

Furthermore, by Theorem 2.3 and $h \sim n^{-\frac{a+2b-1}{4(a+2b)}}$, one has

$$\begin{aligned}
\|\hat{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_0\| &= \|\hat{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h^* + \boldsymbol{\gamma}_h^* - \boldsymbol{\gamma}_0\| \leq \|\hat{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h^*\| + \|\boldsymbol{\gamma}_h^* - \boldsymbol{\gamma}_0\|\\
&= O_p(\delta_n) + O(h^2) = O_p(\delta_n) + O_p(\delta_n) = O_p(\delta_n).
\end{aligned}$$

Note that

$$\begin{aligned}
\|\hat{\beta}_h(t) - \beta_0(t)\|^2 &= \left\|\sum_{j=1}^m \hat{\gamma}_j\hat{\phi}_j - \sum_{j=1}^\infty \gamma_{0j}\phi_j\right\|^2 \leq 2\left\|\sum_{j=1}^m \hat{\gamma}_j\hat{\phi}_j - \sum_{j=1}^m \gamma_{0j}\phi_j\right\|^2 + 2\left\|\sum_{j=m+1}^\infty \gamma_{0j}\phi_j\right\|^2\\
&\leq 4\left\|\sum_{j=1}^m(\hat{\gamma}_j - \gamma_{0j})\hat{\phi}_j\right\|^2 + 4\left\|\gamma_{0j}(\hat{\phi}_j - \phi_j)\right\|^2 + 2\left\|\sum_{j=m+1}^\infty \gamma_{0j}\phi_j\right\|^2\\
&\triangleq 4D_1 + 4D_2 + 2D_3.
\end{aligned}$$

According to Equation (2.15), Condition 2.4, the orthogonality of $\hat{\phi}_j$ and $\|\hat{\phi}_j - \phi_j\|^2 = O_p(n^{-1}j^2)$, we have

$$\begin{aligned}
D_1 &= \left\|\sum_{j=1}^m(\hat{\gamma}_j - \gamma_{0j})\hat{\phi}_j\right\|^2 \leq \left|\sum_{j=1}^m(\hat{\gamma}_j - \gamma_{0j})\right|^2 = \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|^2 = O_p(\delta_n^2),\\
D_2 &= \left\|\gamma_{0j}(\hat{\phi}_j - \phi_j)\right\|^2 \leq m\sum_{j=1}^m\left\|\hat{\phi}_j - \phi_j\right\|^2\gamma_{0j}^2 \leq \frac{m}{n}O_p\left(\sum_{j=1}^m j^2\gamma_{0j}^2\right)\\
&= O_p\left(\frac{m}{n}\sum_{j=1}^m j^{2-2b}\right) = O_p\left(\frac{m}{n}\right) = o_p(n^{-\frac{2b-1}{a+2b}}),\\
D_3 &= \left\|\sum_{j=m+1}^\infty \gamma_{0j}\phi_j\right\|^2 \leq c\sum_{j=m+1}^\infty j^{-2b} = O\left(n^{-\frac{2b-1}{a+2b}}\right).
\end{aligned}$$

Taking these bounds together, we can finish the proof of Theorem 2.3 immediately. □

Define $\mathcal{S} = \{(Y_i, X_i(\cdot)) : 1 \le i \le n\}$. In the following, for a new pair of predictor variables $(Y_{n+1}, X_{n+1}(\cdot))$ taking from the same population as the data and independent of the data, we shall derive the convergence rate of the mean squared prediction error (MSPE) given by

$$\text{MSPE} = \mathrm{E}\Big(\Big[\Big(\int_0^1 X_{n+1}(t)\hat{\beta}_h(t)dt\Big) - \Big(\int_0^1 X_{n+1}(t)\beta_0(t)dt\Big)\Big]^2\Big|\mathcal{S}\Big).$$

Theorem 2.4. *Form Theorem 2.3 (2) and Conditions 2.4, we have*

$$MSPE = O_p\Big(n^{-\frac{a+2b-1}{a+2b}}\Big).$$

Proof. According to the definition of the MSPE, we have

$$\begin{aligned}\text{MSPE} \le & 3\sum_{j=1}^{m}(\hat{\gamma}_j - \gamma_{0j})^2\lambda_j + 3c\left\|\sum_{j=1}^{m}\hat{\gamma}_j(\phi_j - \hat{\phi}_j)\right\|^2 + 3\sum_{j=m+1}^{\infty}\gamma_{0j}^2\lambda_j \\ \triangleq & 3E_1 + 3cE_2 + 3E_3.\end{aligned} \tag{2.16}$$

In accordance with the preceding proofs and Condition 2.4(a) yields $E_1 \le \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|^2 = O_p(\frac{m}{n})$. As for E_2, based on the triangle inequality and Hölder inequality, we know

$$\begin{aligned}E_2 = & \left\|\sum_{j=1}^{m}\hat{\gamma}_j(\phi_j - \hat{\phi}_j)\right\|^2 = \left\|\sum_{j=1}^{m}\gamma_{0j}(\phi_j - \hat{\phi}_j) + (\hat{\gamma}_j - \gamma_{0j})(\phi_j - \hat{\phi}_j)\right\|^2 \\ \le & 2m\sum_{j=1}^{m}\gamma_{0j}\left\|\phi_j - \hat{\phi}_j\right\|^2 + 2\left\|\hat{\gamma}_j - \gamma_j\right\|^2\left\|\sum_{j=1}^{m}(\phi_j - \hat{\phi}_j)^2\right\| = O_p\left(\frac{m}{n}\right) + O_p\left(\frac{m}{n}\right)O_p\left(\frac{m^3}{n}\right) \\ = & O_p\left(\frac{m}{n}\right).\end{aligned}$$

Furthermore, using Condition 2.4, a simple calculation yields $E_3 = O\Big(m^{-(a+2b-1)}\Big)$. Taking these bounds together with (2.16), we have $\text{MSPE} = O_p\Big(n^{-\frac{a+2b-1}{a+2b}}\Big)$. Then, Theorem 2.4 is proven. □

Remark 2.5. Theorem 2.3 shows consistency of estimators, which is similar to Theorem 4.2 in [31]. Theorem 2.4 gives the convergence rate of the prediction error. In particular,r, prediction $\hat{Y}_{n+1}$ can achieve a faster convergence rate than $\hat{\beta}_h(t)$. The reason behind this is that the integration operation, in computing $\int_0^1 X_{n+1}(t)\hat{\beta}_h(t)dt$ from $\hat{\beta}_h(t)$, provides additional smoothness no matter what level of smoothness is used in constructing $\hat{\beta}_h(t)$. Thus, $\int_0^1 X_{n+1}(t)\hat{\beta}_h(t)dt$ usually becomes oversmoothed when we smooth $\hat{\beta}_h(t)$ optimally for estimating $\beta(t)$. As a result, the construction of $\hat{\beta}_h(t)$, as an initial step in estimating $\int_0^1 X_{n+1}(t)\hat{\beta}_h(t)dt$, should involve significant under-smoothing relative to the amount of smoothing used to estimate $\beta(t)$ itself. This undersmoothing enables $\int_0^1 X_{n+1}(t)\hat{\beta}_h(t)dt$ to be estimated $n^{-\frac{a+2b-1}{a+2b}}$ consistently, even though $\hat{\beta}_h(t)$ itself cannot be estimated at such a fast rate in [33].

3. Algorithm

In this section, we employ the majorize-minimization (MM) principle to derive an iterative algorithm to solve (2.7). We first provide a brief overview of the MM algorithm [34]. Considering the minimization of a general smooth function $f(\boldsymbol{\beta})$, given an estimate $\hat{\boldsymbol{\beta}}^{k-1}$ at the kth iteration, the MM algorithm majorizes $f(\boldsymbol{\beta})$ with a properly constructed function

$g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{k-1})$ that satisfies the property $f(\hat{\boldsymbol{\beta}}^k) \le g(\hat{\boldsymbol{\beta}}^k|\hat{\boldsymbol{\beta}}^{k-1})$, where $\hat{\boldsymbol{\beta}}^k = \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{k-1})$. This ensures the decrease of the objective function after each step, i.e., $f(\hat{\boldsymbol{\beta}}^k) \le f(\hat{\boldsymbol{\beta}}^{k-1})$.

We majorize $\hat{Q}_h(\boldsymbol{\gamma})$ given $\hat{\boldsymbol{\gamma}}_h^{k-1}$ by constructing a quadratic function of the form

$$G(\boldsymbol{\gamma}^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1}) = \hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^{k-1}) + \nabla\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^{k-1})(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_h^{k-1}) + \frac{\varphi_k}{2}\|\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_h^{k-1}\|_2^2,$$

where $\varphi_k > 0$ is a quadratic parameter (to be determined) at the kth iteration. Then, define the kth iterate $\hat{\boldsymbol{\gamma}}_h^k$ as the solution to

$$\text{minimize}_{\boldsymbol{\gamma}\in\mathbb{R}^m} G(\boldsymbol{\gamma}_h^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1}). \tag{3.1}$$

To ensure the descent of the objective function in (2.7) at each iteration, the parameter $\varphi_k > 0$ must be sufficiently large such that $\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^k) \le G(\hat{\boldsymbol{\gamma}}_h^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1})$. Consequently,

$$\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^k) \le G(\hat{\boldsymbol{\gamma}}_h^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1}) \le G(\hat{\boldsymbol{\gamma}}_h^{k-1}|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1}) \le \hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^{k-1}),$$

where the second inequality is due to the fact that $\hat{\boldsymbol{\gamma}}_h^k$ is a minimizer of (3.1). In practice, we choose φ_k by starting from a small value $\varphi_0 = 0.01$ and successively inflate it by a factor $\alpha = 1.2$ until the majorization requirement $\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^k) \le G(\hat{\boldsymbol{\gamma}}_h^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1})$ is met at each iteration of the MM algorithm.

By the first-order optimization condition, $\hat{\boldsymbol{\gamma}}_h^k$ satisfies $\mathbf{0} \in \nabla\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^{k-1}) + \varphi_k(\hat{\boldsymbol{\gamma}}_h^k - \hat{\boldsymbol{\gamma}}_h^{k-1})$. Its update takes a simple form $\hat{\boldsymbol{\gamma}}_h^k = \hat{\boldsymbol{\gamma}}_h^{k-1} - \varphi_k^{-1}\nabla\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^{k-1})$. Detailed update rules of $\hat{\boldsymbol{\gamma}}_h^k$ are summarized in the following Algorithm 1.

Algorithm 1 The Algorithm for Solving (2.7).

Input: Kernel function $K(\cdot)$, bandwidth h, turning parameter m, inflation factor $\alpha = 1.2$, and convergence criterion υ.

1. $\hat{\boldsymbol{\gamma}}_h^k = \mathbf{0}, \varphi_0 = 0.01$. Repeat the following steps until the stopping criterion $\|\hat{\boldsymbol{\gamma}}_h^k - \hat{\boldsymbol{\gamma}}_h^{k-1}\| \le \upsilon$ is met, where $\hat{\boldsymbol{\gamma}}_h^k$ is the kth iterate, $\upsilon = 10^{-6}$.
2. Set $\varphi_k \leftarrow \max\{\varphi_0, \frac{\varphi_k}{\alpha}\}$.
3. If $\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^k) \le G(\hat{\boldsymbol{\gamma}}_h^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1})$, set $\varphi_k = \alpha\varphi_k$.
4. Until $\hat{Q}_h(\hat{\boldsymbol{\gamma}}_h^k) \le G(\hat{\boldsymbol{\gamma}}_h^k|\varphi_k, \hat{\boldsymbol{\gamma}}_h^{k-1})$.

Output: the updated parameter $\hat{\boldsymbol{\gamma}}_h^k$.

4. Simulation Study

In our simulation studies, we systematically evaluate the performance of our proposed SLAD method against five competing approaches: (1) Ordinary least squares (OLS), (2) LAD, (3) functional Bayesian method (FB) in [29], which implemented using the bayesQR(0.5) package in R with 200 posterior draws (ndraw=200) and default prior specifications, (4) functional B-spline method (FSp) in [27], (5) functional penalized smoothing spline method (FPSp) in [28], where the penalty parameter $\lambda = 1$ was chosen following common practice in penalty-spline literature [35], with sensitivity checks confirming robustness to $\lambda \in [0.1, 10]$.

Our simulation design considers three sample sizes (n=200, 400, 600) with 500 replications for each scenario. Data are generated from the following model:

$$Y = \int_0^1 X(t)\beta(t)dt + \sigma(X)\{\epsilon - \text{median}(\epsilon)\},$$

where $X(t) = \sum_{j=1}^{50}\xi_j\phi_j(t)$, $\xi_j \sim N(0, \lambda_j)$, variance $\lambda_j = ((j-0.5)\pi)^{-2}$, $\phi_j(t) = \sqrt{2}\sin((j-0.5)\pi t)$, functional coefficients $\beta(t) = \sqrt{2}\sin(\frac{\pi t}{2}) + 3\sqrt{2}\sin(\frac{3\pi t}{2})$. $\epsilon - \text{median}(\epsilon)$ such that the median error is 0 for identification purposes. We consider four different distributions for ϵ as follows:

Table 1. MISE(IV) (×100) under homoscedastic error for case (A).

Error	n	Methods					
		OLS	LAD	FSLAD	FB	FSp	FPSp
$N(0,1)$	200	21.26(21.16)	27.41(27.34)	24.09(24.02)	25.96(24.36)	1001.59(27.54)	975.74(1.74)
	400	10.46(10.39)	13.81(13.74)	12.54(12.47)	13.45(13.20)	986.19(12.01)	975.23(1.07)
	600	7.72(7.69)	10.49(10.44)	9.52(9.47)	10.31(10.06)	983.55(9.38)	975.09(0.94)
$t(3)$	200	45.72(45.63)	34.46(34.33)	29.97(29.88)	30.64(30.31)	1019.00(44.93)	976.33(2.32)
	400	22.26(22.25)	15.70(15.67)	13.81(13.79)	14.08(13.88)	992.66(18.48)	975.54(1.42)
	600	15.58(15.57)	11.74(11.73)	10.27(10.26)	10.83(10.66)	987.92(13.82)	975.29(1.21)
MN	200	54.44(54.14)	31.69(31.54)	27.16(27.07)	28.15(27.83)	1013.82(39.57)	976.26(2.13)
	400	29.09(28.84)	15.78(15.66)	14.02(13.94)	14.80(14.38)	990.84(16.60)	975.52(1.30)
	600	20.21(20.20)	11.62(11.55)	10.46(10.40)	10.77(10.43)	986.52(12.35)	975.29(1.14)
Cau(0,1.5)	200	$>10^5(>10^5)$	63.41(63.39)	60.16(60.12)	60.46(60.40)	1174.47(200.22)	980.18(6.25)
	400	$>10^5(>10^5)$	30.35(30.30)	28.13(28.09)	28.69(28.54)	1040.02(65.96)	977.52(3.53)
	600	$>10^5(>10^5)$	25.50(25.50)	24.68(24.67)	24.65(24.58)	1022.96(48.89)	976.98(2.98)

(i) Standard normal distribution $N(0,1)$;
(ii) t distribution with the degrees of freedom 3, $t(3)$;
(iii) Mixture of normals (MN) $0.9N(-1,1)+0.1N(1,5)$;
(iv) Cauchy distribution with location parameter of 0 and a scale parameter of 1.5, Cau(0,1.5).

In addition, the error term $\sigma(X)$ is generated from one of the following three cases:
(A) (Homoscedastic) $\sigma(X)=1$;
(B) (Linear heteroscedastic) $\sigma(X)=0.37\int_0^1 X(t)\beta(t)dt+1$;
(C) (Quadratic heteroscedastic) $\sigma(X)=0.37\left(1+\left(\int_0^1 X(t)\beta(t)dt-1\right)^2\right)$.

We implement the SLAD method using the triangular kernel, and the smoothing bandwidth is the same as [25], that is, $h=\max\{0.05,0.5\{\log(m)/n\}^{1/4}\}$. We note that our numerical experiments are rather insensitive to the choice of h provided that it is in a reasonable range (neither too small nor too large). In addition, we use the integrated variance (IV) and mean integrated squared error (MISE) to assess the performance of the estimation for slope function $\beta(\cdot)$:

$$\mathrm{IV}(\hat{\alpha}(t))=\frac{1}{d}\sum_{k=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{hi}(t_k)-\frac{1}{n}\sum_{i=1}^{n}\hat{\beta}_i(t_k)\right)^2,\ \mathrm{MISE}=\frac{1}{d}\sum_{k=1}^{d}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{hi}(t_k)-\beta_i(t_k)\right)^2,$$

where $\{t_k:\ k=1,2,\ldots,d\}$ are grid points chosen to be equally spaced in the domains of function $\beta(\cdot)$. In our simulation, $d=100$ is used. For each scenario, the tuning parameter m is determined by CPV criterion as described in Section 2.

Tables 1–3 show IVs (in parentheses) and MISEs of $\hat{\beta}_h(\cdot)$ under different errors. The simulation studies yield the following key findings: (i) as expected, both MISEs and IVs decrease and the performance of the estimation improves as the sample size n increases from 200 to 600. (ii) under ideal Gaussian homoscedastic errors, OLS achieves optimal performance, while the proposed SLAD method closely approximates OLS and significantly outperforms other methods; (iii) for heavy-tailed, outlier and the heteroscedastic

Table 2. MISE(IV) (×100) under linear heteroscedastic error for case (B).

error	n	Methods					
		OLS	LAD	FSLAD	FB	FSp	FPSp
$N(0,1)$	200	21.77(21.74)	20.42(20.40)	19.02(19.01)	19.65(19.57)	1000.11(26.12)	975.50(1.53)
	400	10.78(10.77)	9.49(9.49)	9.04(9.04)	9.48(9.40)	985.22(11.13)	974.97(0.95)
	600	8.67(8.66)	7.92(7.92)	7.66(7.66)	7.86(7.80)	982.59(8.49)	974.90(0.81)
$t(3)$	200	61.44(61.42)	24.80(24.65)	23.86(23.73)	24.65(24.24)	1017.60(43.30)	976.27(2.09)
	400	26.88(26.86)	10.97(10.95)	10.84(10.83)	11.38(11.23)	990.96(16.84)	975.30(1.24)
	600	22.24(22.14)	9.96(9.95)	9.61(9.59)	9.93(9.82)	986.61(12.44)	975.19(1.07)
MN	200	65.18(62.89)	22.61(22.50)	21.56(21.47)	22.21(21.92)	1013.37(38.98)	976.14(1.95)
	400	37.38(33.57)	11.94(11.93)	11.21(11.20)	11.68(11.57)	989.19(15.07)	975.27(1.18)
	600	28.01(25.98)	9.29(9.15)	8.96(8.81)	9.25(9.13)	985.91(11.62)	975.28(0.99)
Cau(0,1.5)	200	$> 10^5(> 10^5)$	61.31(61.30)	59.73(59.71)	60.30(60.10)	1215.12(240.80)	981.00(7.06)
	400	$> 10^5(> 10^5)$	30.01(29.96)	29.53(29.49)	30.21(30.02)	1054.53(80.29)	978.02(3.92)
	600	$> 10^5(> 10^5)$	21.50(21.49)	20.54(20.53)	21.14(21.07)	1031.56(57.38)	977.29(3.28)

Table 3. MISE(IV) (×100) under quadratic heteroscedastic error for case (C).

error	n	Methods					
		OLS	LAD	FSLAD	FB	FSp	FPSp
$N(0,1)$	200	45.45(45.33)	23.72(23.61)	21.63(21.50)	21.78(21.39)	997.71(23.47)	975.38(1.22)
	400	20.49(20.45)	11.58(11.55)	10.71(10.69)	11.27(11.13)	982.75(8.66)	974.79(0.73)
	600	16.13(16.11)	8.87(8.86)	8.28(8.26)	8.78(8.65)	980.46(6.35)	974.69(0.60)
$t(3)$	200	129.22(128.71)	28.18(27.96)	26.18(25.97)	26.33(25.75)	1010.80(36.57)	975.82(1.65)
	400	59.01(58.98)	13.60(13.57)	13.17(13.16)	13.36(13.20)	987.24(13.17)	975.03(0.97)
	600	45.71(45.63)	10.44(10.43)	10.01(9.99)	10.28(10.14)	983.90(9.78)	974.90(0.82)
MN	200	121.53(109.99)	27.14(27.10)	25.86(25.82)	25.92(25.90)	1007.27(33.38)	975.48(1.62)
	400	69.46(55.65)	13.99(13.93)	13.14(13.10)	13.61(13.39)	986.10(12.02)	974.96(0.91)
	600	59.69(45.35)	10.26(10.20)	9.52(9.48)	9.93(9.72)	982.59(8.51)	974.82(0.75)
Cau(0,1.5)	200	$> 10^5(> 10^5)$	100.84(100.78)	96.21(96.12)	96.62(96.58)	1290.13(316.20)	981.48(7.73)
	400	$> 10^5(> 10^5)$	45.26(45.23)	44.18(44.16)	44.37(44.30)	1064.62(90.59)	977.95(4.10)
	600	$> 10^5(> 10^5)$	35.76(35.65)	34.41(34.29)	34.48(34.42)	1038.78(64.84)	977.11(3.27)

scenarios cases, SLAD exhibits superior robustness; (iv) for extreme heavy tails conditions (Cauchy errors), SLAD remains valid, while OLS fails catastrophically; (v) the FB method shows comparable robustness but requires intensive computation, whereas spline-based approaches (FSp/FPSp) perform poorly due to basis misalignment.

The simulation results reveal the performance differences of various functional regression methods under different error distributions. For FPCA-based methods, OLS performs excellently under homoscedastic normal errors but deteriorates sharply due to sensitivity to outliers in heavy-tailed distributions. In contrast, the LAD method demonstrates superior

robustness through median-based estimation. However, the FB exhibits instability, probably due to sensitivity to prior distributions and limitations in finite samples. However, spline-based methods (FSp and FPSp) generally underperform, particularly with smaller sample sizes, as their smoothing strategies relying on knot selection struggle to adapt to complex error structures. Notably, although the integral variances (IVs) of spline methods are smaller, their overall accuracy in this simulation setup still lags behind FPCA-based methods.

In general, FPCA-based methods outperform spline-based methods in this study, with SLAD proving most reliable for handling heavy-tailed or heteroscedastic errors. Although the OLS remains optimal under Gaussian homoscedastic conditions, its vulnerability in other scenarios underscores the importance of robust alternatives. The instability of FB and the poor performance of spline methods highlight the need to weigh specific scenarios when selecting methods, emphasizing that error distribution characteristics should be prioritized. These findings align with theoretical expectations and position SLAD as a versatile default choice, bridging the robustness-efficiency trade-off in functional regression.

5. Application

To illustrate the utility of the proposed method, we apply our proposed procedure to analyze Berkeley growth data and Capital Bike Share data, respectively. For each dataset, the cut-off parameter m is determined by CPV criterion as described in Section 2. The bandwidth $h = \max\{0.05, 0.5\{\log(m)/n\}^{1/4}\}$. Similar to [33], each dataset is randomly divided into two subsamples: the training sample, $I_1 = \{(X_i, Y_i), |I_1|\}$, where $|I_1|$ denotes the cardinality of I_1, and the remaining is the test sample, $I_2 = \{(X_i, Y_i), |I_2|)\}$. The training sample is used to estimate the parameters, and the test sample is employed to verify the quality of predictions. This random splitting is repeated $N(N = 50, 100, 200)$ times. Two types of MSPE, namely, the mean square prediction error (MSPE) and median square prediction error (MedSPE), are considered based on the test set (see, e.g., [36]) as follows:

$$\text{MSPE} = \frac{1}{n_{\mathcal{J}}} \sum_{j \in \mathcal{J}} \frac{(Y_j - \hat{Y}_j)^2}{s^2_{\mathcal{J}}} \quad \text{and} \quad \text{MedSPE} = \frac{\text{median}(Y_j - \hat{Y}_j)^2}{s^2_{\mathcal{J}}}, \ (j \in \mathcal{J}),$$

where $\mathcal{J}$ contains the indices of the observations in the test set, $n_{\mathcal{J}}$ denotes its size, and $s_{\mathcal{J}} = \text{MAD}_{j \in \mathcal{J}}(Y_j)$ with MAD is the abbreviation of "median absolute deviation".

In our empirical analysis, we maintain the exact simulation configurations while comparing our method against five benchmarks: OLS, LAD, FB, FSp, and FPSp (with $\lambda = 1$), ensuring identical functional representations across all methods.

5.1. Application to Berkeley Growth data

In this subsection, we apply our proposed test procedure to analyze Berkeley growth data in [37], which is available in the R package "fda". The heights of 54 girls and 39 boys aged 1 to 18 in a set of 31 ages were collected. The data collection protocol involved four measurements during the first year, annual measurements from ages 2 to 8, followed by biannual measurements thereafter. Previous studies [37, 38] have consistently demonstrated significant gender differences in growth patterns, particularly during adolescence (ages 13-18), motivating our sex-specific modeling approach. Our analysis focuses on predicting adult height Y from growth trajectories $X(t)$ over varying intervals $[1, T]$, where $t \in [1, T]$ represents age and T ranges from 11 to 17 years. We employ functional linear models with $m = 3$ principal components, which explain approximately 99% of the variance of height, while accounting for the documented differences in growth rates between boys and girls. To validate our models, we performed Shapiro-Wilk tests to assess residual normality for each T and established training sets of $|I_1| = 36$ girls and $|I_1| = 26$

boys, with complete performance results in $N \in \{50, 100, 200\}$ and prediction intervals $T = 11, \cdots, 17$ presented in Tables 4-5.

From Table 4, we can see that the p-values of Shapiro-Wilk test are all less than 0.05 from $T = 11$ to $T = 17$, thereby indicating the nonnormal distribution of the residuals. As expected, Table 4 shows that SLAD has the best prediction performance than the other methods, as its MSPE and MedSPE is the smallest, which indicates that our method fits girls' data better than the LAD, OLS, LAD, FB, FSp and FPSp methods. These results reconfirm that the proposed method can again give a robust and efficient estimator. Table 5 shows that the p-values of Shapiro-Wilk test are all more than 0.05 from $T = 11$ to $T = 17$, thereby the residuals tend to be normal distribution. The results of Table 5 show that the OLS estimator performs the best for normal errors, and the proposed SLAD estimator performs comparably with the OLS estimator and clearly better than the remaining methods.

Our analysis of the Berkeley Growth data reveals critical insights into the performance of FLM in predicting adult height. The results demonstrate that the accuracy of the prediction is strongly influenced by biological sex and developmental timing, reflecting known differences in growth trajectory. For girls, the smaller prediction errors occur when using pre-pubertal data (T=11-13), with error peaking during mid-puberty ($T = 15$). This aligns with Tanner's growth phase theory, as girls experience their height velocity peak earlier ($10 - 12$ years) with rapid nonlinear changes that challenge modeling. In contrast, boys maintain stable low errors until $T = 16$, consistent with their prolonged adolescent growth spurt ($12 - 18$ years). The proposed SLAD method outperforms alternatives by adaptively handling these nonlinearities. This advantage comes from the ability of FPCA to represent phase variation [1], where the first three principal components explain 99% of variance by capturing: (i) baseline growth; (ii) pubertal acceleration; and (iii) individual timing differences. Although FB approaches show comparable precision for boys, their performance degrades with atypical growth patterns due to rigid priors. Future work should explore dynamic penalization to further improve pubertal-phase predictions.

5.2. Application to Capital Bike Dhare data

Renting bicycles has become increasingly popular in recent years because it is considered a more economical and environmentally friendly alternative to owning bicycles. Ensuring sufficient bike supply is critical for a successful business. We applied the proposed method to the analysis of bike rental data. The dataset contains information on counts of casual bike rentals during the period from January 1, 2011 to December 31, 2012, or a total of 105 weeks from Capital Bikeshare System (CBS), Washington D.C., USA, which is available at http://capitalbikeshare.com/system-data.

Considering bike rentals have different dynamics on weekends from that on weekdays, we restricted our analysis to rentals on Saturdays, in which the demand for casual bike rentals is higher compared with the demand for weekdays'bike rentals. We aimed to examine how Saturday rentals were related to temperature. Understanding the nature of this association can help predict the demand for casual rentals based on weather forecast. Following the analysis of [39] and [19], we consider the daily count of bike rentals as a scalar response variable and perform a logarithmic transformation $Y \to \log(Y + 1)$ to remove skewness. We considered the centered hourly temperature curve as functional covariate $X(t)$. We chose $|I_1| = 55$ in these data. The hourly temperature included a small amount of missingness. We used the functional principal component analysis method and "refund" package ("fpca.sc" function) in R to impute the missing data before applying the center/scaling transformation. Here, $m = 3$ is selected to explain approximately 99% of the variance of the temperature. The OLS method, is initially used to fit the model, and the fitting results of the residuals are shown in Figure 1.

Table 4. Results (×100) of different methods for girls in Berkeley growth data analysis.

N	Criterions	Methods	T						
			T=11	T=12	T=13	T=14	T=15	T=16	T=17
	p-values		2.885	2.485	3.147	2.669	0.060	0.300	2.210
50	MSPE	OLS	2.529	0.470	0.781	4.186	13.168	8.045	6.377
		LAD	1.752	0.418	0.748	4.180	13.009	7.946	6.079
		FSLAD	1.723	0.410	0.728	3.903	12.853	7.902	6.064
		FB	1.635	0.408	0.731	4.067	12.953	7.942	6.039
		FSp	84.448	15.953	21.856	75.215	81.233	82.580	96.911
		FPSp	74.867	14.523	11.598	52.819	43.143	62.636	78.571
	MedSPE	OLS	2.504	0.413	0.722	4.055	13.093	8.010	6.283
		LAD	1.712	0.365	0.697	4.047	12.932	7.900	5.977
		FSLAD	1.676	0.354	0.675	3.772	12.781	7.863	5.964
		FB	1.576	0.356	0.678	3.972	12.855	7.864	5.971
		FSp	89.931	15.337	21.789	74.622	81.086	72.479	96.965
		FPSp	74.092	14.180	11.589	52.097	43.026	52.553	78.604
100	MSPE	OLS	1.161	0.388	0.887	3.704	8.771	7.937	5.685
		LAD	1.186	0.383	0.830	3.609	8.727	7.892	5.660
		FSLAD	1.172	0.370	0.816	3.572	8.706	7.854	5.624
		FB	1.173	0.382	0.821	3.598	8.714	7.933	5.698
		FSp	60.156	13.530	19.772	43.246	92.017	27.578	72.943
		FPSp	53.646	12.548	15.748	41.275	54.460	22.521	70.313
	MedSPE	OLS	1.143	0.340	0.824	3.607	8.700	7.860	5.591
		LAD	1.163	0.334	0.770	3.515	8.654	7.811	5.568
		FSLAD	1.154	0.322	0.754	3.480	8.633	7.772	5.529
		FB	1.096	0.330	0.768	3.502	8.642	7.779	5.560
		FSp	60.084	13.161	25.215	42.867	91.734	29.645	72.070
		FPSp	53.597	12.223	15.717	41.988	54.292	26.737	69.433
200	MSPE	OLS	2.870	0.489	0.660	4.573	9.768	6.627	5.982
		LAD	2.090	0.467	0.638	4.545	9.682	6.460	5.730
		FSLAD	1.973	0.460	0.616	4.527	9.634	6.456	5.724
		FB	1.852	0.463	0.630	4.530	9.642	6.537	5.728
		FSp	67.689	15.915	18.805	56.202	89.039	92.798	70.964
		FPSp	53.877	12.589	12.724	54.186	51.137	72.385	68.380
	MedSPE	OLS	2.802	0.431	0.593	4.459	9.672	6.513	5.906
		LAD	2.018	0.413	0.571	4.429	9.589	6.344	5.653
		FSLAD	1.907	0.404	0.551	4.409	9.538	6.341	5.647
		FB	1.778	0.405	0.570	4.411	9.584	6.448	5.650
		FSp	67.572	15.411	18.213	54.248	88.836	82.872	70.010
		FPSp	53.784	11.956	12.417	49.578	51.055	71.385	67.412

Figure 1 (a) shows the diagnostic plot of residuals and Figure 1 (b) shows the density of the residuals of the FLM adjustment. Apparently, the data exhibit a heavy-tailed distribution and three outliers are present. In addition, a Shapiro-Wilk test is performed to assess the normality of the residuals. The corresponding p-value is 1.6×10^{-3}, thereby reconfirming the significant deviation from Gaussian assumptions and justifying the need

Table 5. Results (×100) of different methods for boys in Berkeley growth data analysis.

N	Criterions	Methods	T						
			T=11	T=12	T=13	T=14	T=15	T=16	T=17
	p-values		19.210	72.800	98.310	98.810	96.460	81.490	96.340
50	MSPE	OLS	1.217	0.724	0.766	0.770	0.706	0.568	1.837
		LAD	1.422	0.828	0.902	1.078	1.056	0.809	2.397
		FSLAD	1.319	0.734	0.861	1.006	0.994	0.806	2.088
		FB	1.327	0.785	0.885	1.072	0.995	0.807	2.143
		FSp	44.200	40.893	11.554	15.373	9.831	6.776	11.259
		FPSp	35.773	18.771	7.250	10.432	7.835	4.250	7.572
	MedSPE	OLS	1.182	0.694	0.705	0.747	0.675	0.542	1.812
		LAD	1.359	0.791	0.849	1.052	1.026	0.776	2.378
		FSLAD	1.263	0.699	0.805	0.981	0.963	0.774	2.063
		FB	1.329	0.747	0.867	1.050	0.975	0.784	2.175
		FSp	44.184	40.790	11.869	15.362	13.995	6.801	11.076
		FPSp	35.778	18.613	7.349	10.920	11.168	4.087	7.572
100	MSPE	OLS	1.031	0.519	0.580	1.067	0.441	0.594	2.723
		LAD	1.368	0.562	0.826	1.246	0.832	0.949	3.169
		FSLAD	1.226	0.534	0.775	1.135	0.713	0.850	3.056
		FB	1.248	0.540	0.792	1.152	0.717	0.860	3.055
		FSp	70.158	50.441	14.135	13.926	15.562	5.927	11.411
		FPSp	47.292	25.210	7.831	8.154	12.578	5.059	8.255
	MedSPE	OLS	0.996	0.485	0.551	1.035	0.414	0.559	2.700
		LAD	1.314	0.525	0.792	1.225	0.808	0.920	3.146
		FSLAD	1.178	0.501	0.749	1.109	0.689	0.823	3.035
		FB	1.219	0.507	0.753	1.150	0.691	0.838	3.051
		FSp	69.758	49.887	14.502	14.183	15.023	5.798	11.614
		FPSp	47.043	24.749	7.975	10.421	12.327	4.920	8.464
200	MSPE	OLS	1.416	0.684	0.519	1.253	0.677	0.710	2.810
		LAD	1.540	0.792	0.732	1.444	1.063	1.103	3.127
		FSLAD	1.374	0.738	0.687	1.356	0.944	1.010	3.045
		FB	1.455	0.730	0.702	1.438	1.065	1.007	3.058
		FSp	60.038	52.824	12.780	10.863	16.973	6.546	11.491
		FPSp	45.176	24.736	7.924	7.390	13.253	5.574	8.038
	MedSPE	OLS	1.375	0.660	0.482	1.228	0.646	0.671	2.770
		LAD	1.492	0.762	0.689	1.420	1.033	1.068	3.089
		FSLAD	1.328	0.707	0.645	1.329	0.916	0.975	3.003
		FB	1.426	0.711	0.681	1.345	1.032	0.988	3.040
		FSp	59.776	52.515	12.491	10.597	15.974	6.373	11.292
		FPSp	45.008	24.375	7.806	8.256	13.353	5.387	7.756

for robust estimation methods like the proposed SLAD approach. Table 6 presents the average MSPEs and MedSPEs of N times repeated operations. The results demonstrate that our proposed SLAD method outperforms all competitors. It is worth pointing out the spline-based methods (including FSp and FPSp) perform rather badly. Figure 2 presents the estimated functional coefficients $\hat{\beta}_h(t)$ obtained from four competing methods: (i) our

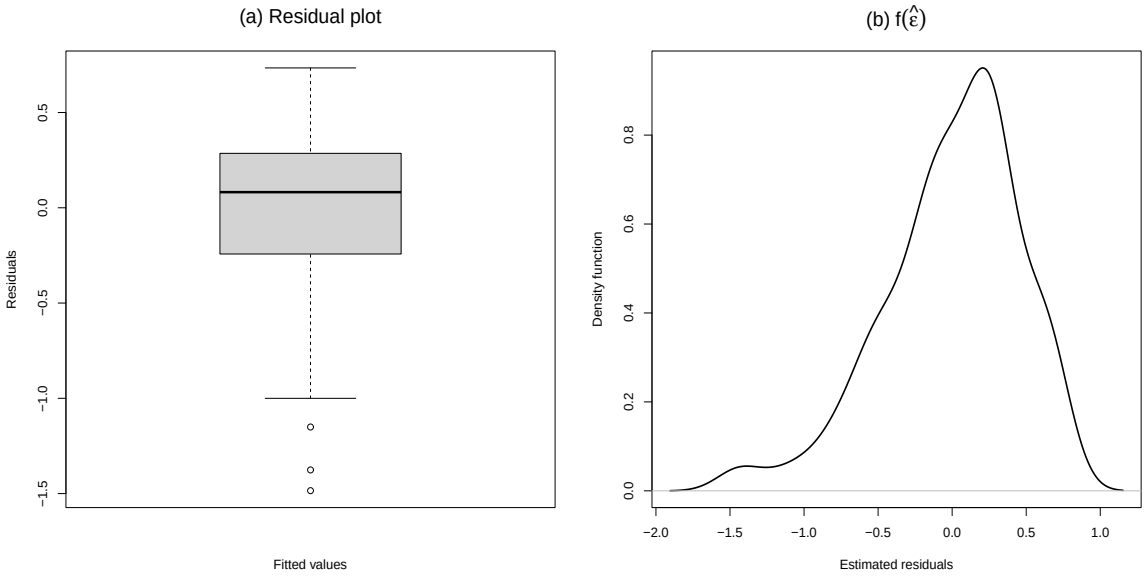

Figure 1. (a) The residual plot with the OLS method; (b) The density of estimated errors with the OLS method.

Table 6. Results (×100) of different methods in bike sharing data analysis.

Criterions	N	Estimation methods					
		OLS	LAD	FSLAD	FB	FSp	FPSp
MSPE	50	0.433	0.431	0.414	0.430	2.917	1.871
	100	0.426	0.424	0.414	0.427	2.868	1.883
	200	0.432	0.428	0.418	0.390	1.987	1.324
MedSPE	50	0.215	0.196	0.195	0.196	2.040	1.343
	100	0.194	0.177	0.176	0.176	1.946	1.328
	200	0.189	0.179	0.178	0.177	2.863	1.852

proposed SLAD estimator (solid line); (ii) FB approach (dashed line); (iii) FSp method (dotted line); and (iv) FPSp method (dotdash line), where the estimated functional coefficients by OLS and LAD methods are similar to FLAD method, and will not be presented. The graph demonstrates significant differences in the performance characteristics of the four estimation methods. The proposed SLAD estimator (solid curve) and FB (yellow curve) show remarkably similar trend patterns, both successfully capturing the fundamental inverted U-shaped relationship between temperature and bike rentals. In general, the effect of temperature on counts of bike rentals is positive, implying an increase in biking with a rise in temperature. Such effect reaches its maximum in midday, particularly from 9:00 am to 15:00 pm, during which the temperature tends to be warm and people are most likely to bike. In early or late hours, the effect of temperature on bike rental decreases because people are less likely to bike during these time periods. This finding coincides with the result of [20] and our intuitive thinking. In contrast, spline-based methods show notable limitations. The FSp (dotted line) produces estimates that are excessively volatile. While the FPSp method (dotdash line) reduces this variability, it introduces substantial bias by oversmoothing important features. Hence, the spline-based methods, particularly the FSp approach, fail to accurately capture the dynamic characteristics of the actual temperature-rental relationship.

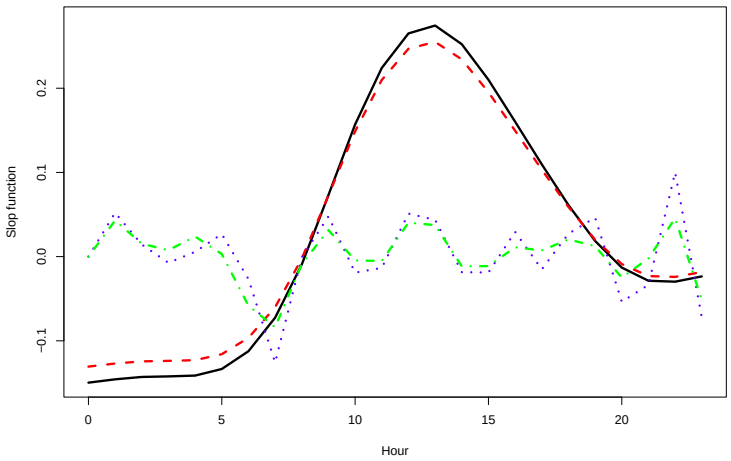

Figure 2. The estimated functional coefficient $\hat{\beta}_h(t)$ with FSLAD (solid line), FB (dashed line), FSp (dotted line), and FPSp (dotdash line) methods.

6. Concluding Remarks

In this study, we have developed a novel FLM SLAD estimator, which successfully addresses the computational challenges of standard LAD estimation while improving the efficiency of the estimation. The key methodological innovations include: (i) constructing a smoothed approximation to the non-differentiable LAD objective function; (ii) employing functional principal component analysis for effective dimension reduction in infinite-dimensional parameter spaces; and (iii) establishing the estimator's asymptotic properties, confirming that the SLAD estimator achieves the same convergence rate as the LAD estimator when the bandwidth approaches zero and deriving convergence rate for prediction error. The proposed method demonstrates superior performance in several important aspects: (i) enhanced robustness against heavy-tailed distributions, outliers, and heteroscedastic errors compared to conventional approaches (OLS, LAD); (ii) maintained estimation efficiency under ideal conditions; (iii) practical advantages over spline-based regularization and Bayesian methods in finite-sample scenarios; (iv) improved predictive performance and interpretability in real-life applications, as evidenced by the Berkeley Growth study and Capital Bike Share dataset analyses. Our analysis of the Berkeley growth data reveals critical insights into the performance of FLM for predicting adult height. The results demonstrate that prediction accuracy is strongly influenced by biological sex and developmental timing, reflecting known growth trajectory differences. In the Capital Bike Share dataset analyses, the estimated temperature effects reveal a consistent inverted U-shaped relationship across methods, with peak influence occurring in mid-afternoon (9:00-15:00 PM). This pattern suggests that riders are most sensitive to temperature variations during typical leisure hours. The functional coefficient plots show that SLAD and FB methods capture this basic shape, while SLAD produces smoother and more stable estimates, particularly at temperature extremes where other methods show erratic fluctuations. From an operational perspective, these findings suggest that bike share systems could optimize their redistribution strategies by: (i) increasing inventory at stations prior to the afternoon temperature peak; (ii) implementing dynamic pricing during high-demand temperature windows; (iii) using SLAD-based predictions to anticipate demand surges. The superior performance of SLAD in this application stems from its ability to: (i) handle the inherent skewness in rental count data; (ii) accommodate

nonlinear temperature effects; (iii) maintain robustness against outliers; (iv) provide stable estimates across different sample sizes. These advantages make SLAD particularly valuable for transportation systems planning and real-time management decisions. Future extensions could incorporate additional weather variables and examine weekday/weekend differences more systematically.

Several promising directions for future research emerge from this work: (i) extension to functional semi-parametric and non-parametric modeling frameworks; (ii) incorporation of penalization techniques for variable selection in high-dimensional functional regression settings; (iii) generalization to more complex data structures, including: dependent functional data, partially observed functional data and multivariate functional data. These extensions would require substantial theoretical development and computational innovation, particularly in addressing the challenges of dependent observations and incomplete functional data. Future work should also investigate optimal bandwidth selection strategies and adaptive smoothing approaches to further enhance the practical utility of the proposed method.

Acknowledgements

We would like to thank editor and two anonymous referees for their careful reading and insightful comments, which led to significant improvements of this paper.

Author contributions. All the co-authors have contributed equally in all aspects of the preparation of this submission.

Conflict of interest statement. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This work is supported by National Natural Science Foundation of China (12071267,12401356), Natural Science Foundation of Shanxi Province (202203021222223), National Statistical Science Research Project of China (2022LY089) and the Natural Science Foundation of Shanxi normal University (JYCJ2022004).

Data availability. Open-access data is used in this research.

References

[1] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer, New York, 2005.

[2] P. Yushkevich, S.M. Pizer and S. Joshi et al., *Intuitive, localized analysis of shape variability. Inf. Process. Med. Imaging*, Springer, Berlin, 402–408, 2001.

[3] N. Locantore, J.S Marron and D.G. Simpson et al., *Robust principal component analysis for functional data*, Test. **8**, 1–73, 1999.

[4] J.O. Ramsay and B.W. Silverman, *Applied functional data analysis: methods and case studies*, Springer, New York, 2002.

[5] F. Ferraty, *Nonparametric functional data analysis*, Springer, New York, 2006.

[6] L. Horváth and P. Kokoszka, *Inference for functional data with applications*, Springer, New York, 2012.

[7] T. Hsing and R. Eubank, *Theoretical foundations of functional data analysis, with an introduction to linear operators*, John Wiley & Sons, Vol. 997, 2015.

[8] P. Kokoszka and M. Reimherr, *Introduction to functional data analysis*, Chapman and Hall/CRC, New York, 2017.

[9] C.M. Crainiceanu, J. Leroux and E. Cui, *Functional Data Analysis with R*, Chapman and Hall/CRC, New York, 2024.

[10] H. Cardot, F. Ferraty and P. Sarda, *Functional linear model*, Stat. Probab. Lett. **45** (1), 11–22, 1999.
[11] T.T. Cai and P. Hall, *Prediction in functional linear regression*, Ann. Stat. **34** (5), 2159–2179, 2006.
[12] P. Hall and J.L. Horowitz, *Methodology and convergence rates for functional linear regression*, Ann. Stat. **35** (1), 70–91, 2007.
[13] A.R. Rao and M. Reimherr, *Nonlinear functional modeling using neural networks*, J. Comput. Graph. Stat. **32** (4), 1248–1257, 2023.
[14] H. Yeon, X. Dai and D.J. Nordman, *Bootstrap inference in functional linear regression models with scalar response*, Bernoulli. **29** (4), 2599–2626, 2023.
[15] I. Kalogridis and S. Van Aelst, *Robust penalized estimators for functional linear regression*, J. Multivar. Anal. **194**, 105104, 2023.
[16] S. Gurer, H.L. Shang and A. Mandal et al., *Locally sparse and robust partial least squares in scalar-on-function regression*, Stat. Comput. **34** (5), 150, 2024.
[17] J. Liu and L. Shi, *Statistical Optimality of Divide and Conquer Kernel-based Functional Linear Regression*, J. Mach. Learn. Res. **25** (155), 1–56, 2024.
[18] Q. Tang and L. Cheng, *Partial functional linear quantile regression*, Sci. China Math. **57**, 2589–2608, 2014.
[19] P. Yu, X. Song and J. Du, *Composite expectile estimation in partial functional linear regression model*, J. Multivar. Anal. 203, 105343, 2024.
[20] R. Ghosal, *Hypothesis Testing and Variable Selection in Functional Concurrent Regression Model*, PhD thesis, NC State Univ, 2019.
[21] J. Martínez, Á. Saavedra and P. J. García-Nieto et al., *Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain)*, Appl. Math. Comput. **241**, 1–10, 2014.
[22] D. Pollard, *Asymptotics for least absolute deviation regression estimators*, Econom. Theory. **7** (2), 186–199, 1991.
[23] R. Koenker and Jr.G. Bassett, *Regression quantiles*, Econometrica. **46** (1), 33–50. 1978.
[24] M. Fernandes, E. Guerre and E. Horta, *Smoothing quantile regressions*, J. Bus. Econ. Stat. **39** (1), 338–357, 2021.
[25] K.M Tan, L. Wang, and W.X. Zhou, *High-dimensional quantile regression: Convolution smoothing and concave regularization*, J. R. Stat. Soc. B. **84** (1), 205–233, 2022.
[26] X. He, X. Pan, K.M. Tan and W.X. Zhou, *Smoothed quantile regression with large-scale inference*, J. Econom. **232** (2), 367–388, 2023.
[27] H. Cardot, C. Crambes and P. Sarda, *Quantile regression when the covariates are functions*, J. Nonparametr. Stat. **17**, 841–856, 2005.
[28] C. Crambes, A. Kneip and P. Sarda, *Smoothing splines estimators for functional linear regression*, Ann. Stat. **37**, 35–72, 2009.
[29] J. Zhang, J. Cao and L. Wang, *Robust Bayesian functional principal component analysis*, Stat. Comput. **35**, 46, 2025.
[30] L. Zhou, B. Wang and H. Zou, *Sparse convoluted rank regression in high dimensions*, J. Amer. Statist. Assoc. **119** (546), 1500–1512, 2024.
[31] Z. Wang, Y. Bai and W.K. Härdle et al., *Smoothed quantile regression for partially functional linear models in high dimensions*, Biom. J. **65** (7), 2200060, 2023.
[32] P. Yu, Z. Zhang and J. Du, *A test of linearity in partial functional linear regression*, Metrika. **79**, 953–969, 2016.
[33] J. Xiao, P. Yu and X. Song et al., *Statistical inference in the partial functional linear expectile regression model*, Sci. China Math. **65** (12), 2601–2630, 2022.
[34] K. Lange, D.R. Hunter and I. Yang, *Optimization transfer using surrogate objective functions*, J. Comput. Graph. Stat. **9** (1), 1–20, 2000.

[35] P. H. C. Eilers and B. D. Marx, *Flexible smoothing with B-splines and penalties*, Statist. Sci. **11**(2), 89–121, 1996.
[36] G. Boente, M. Salibian-Barrera and P. Vena, *Robust estimation for semi-functional linear regression models*, Comput. Stat. Data Anal. **152**, 107041, 2020.
[37] R. Tuddenham and M. Snyder, *Physical growth of California boys and girls from birth to eighteen years*, Univ. Calif. Publ. Child Dev. **1** (2), 183–364, 1954.
[38] J. M. Tanner, *Growth at Adolescence*, Blackwell Scientific Publications, Oxford, 1962.
[39] J.S. Kim, A.M. Staicu and A. Maity et al., *Additive function-on-function regression*, J. Comput. Graph. Stat. **27** (1), 234–244, 2018.