

Research Article

Jun Sun* and Wanrong Liu

Robust estimation for varying coefficient partially functional linear regression models based on exponential squared loss function

<https://doi.org/10.1515/math-2022-0501>

received April 28, 2021; accepted September 11, 2022

Abstract: In this article, we present a new robust estimation procedure based on the exponential squared loss function for varying coefficient partially functional linear regression models, where the slope function and nonparametric coefficients are approximated by functional principal component basis functions and B splines, respectively. Under some mild conditions, the convergence rates of the resulted estimators are obtained. Simulation studies indicate that our proposed method can achieve robustness against outliers or heavy-tail error distributions and perform no worse than the popular least-squares estimation method for the normal error case. Finally, a real data example is used to illustrate the application of the proposed method.

Keywords: exponential squared loss function, varying coefficient partially functional linear regression models, robustness, convergence rates

MSC 2020: 62G05, 62G35

1 Introduction

Varying coefficient models have become an important research topic in modern statistics and have been frequently applied for addressing problems in various scientific areas, such as economics, biomedical science, finance, medicine, and engineering. One superior characteristic that makes the varying coefficient models attractive is that they have the ability to explore the dynamic pattern by allowing coefficients to vary over the so-called index variable. Moreover, varying coefficient models retain the virtues of both parametric and nonparametric models; therefore, the possible modeling biases can be reduced and the curse of dimensionality can be avoided. Recently, much effort has been devoted to studying the methodological, theoretical, and applied sides around varying coefficient models, and we refer, for instance, to [1–6].

Nowadays, technological innovations allow us to create and store large-scale data sets. In many cases, these data sets can be viewed as functions of time or spatial points or some other continua. Such type of data is termed functional data. In recent years, functional data analysis has attracted much attention among statisticians and practitioners. As the fundamental tool and mathematical framework for functional data analysis, the functional linear model (FLM), which characterizes the relationship between a scalar response and a functional predictor with a linear operator, is seeing a rise in popularity in practice. For a

* **Corresponding author: Jun Sun**, College of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, Anhui, 233030, P. R. China; Key Laboratory of Computing and Stochastic Mathematics (Ministry of Education), School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, P. R. China, e-mail: czhsun@126.com

Wanrong Liu: Key Laboratory of Computing and Stochastic Mathematics (Ministry of Education), School of Mathematics and Statistics, Hunan Normal University, Changsha, Hunan 410081, P. R. China; College of General Education, Hunan University of Information Technology, Changsha, Hunan 410151, P. R. China, e-mail: liuwr08@163.com

systematic coverage of statistical methods and inferences of FLM, see [7–13] and others. It is well known that the FLM model is useful. But in many practical situations, the simple relationship between the response variable and functional covariate may not be sufficient. To improve the power of prediction and interpretation of FLM, several extensions of FLM have been proposed for the analysis of mixed data with complex structures. Recently, Peng et al. [14] incorporated some real-valued covariates in FLM and proposed a mixed data model, namely, a varying coefficient partially functional linear regression model (VCPFLRM). The VCPFLRM takes the form

$$Y = Z^T \beta(U) + \int_{\mathcal{T}} \alpha(t) X(t) dt + \varepsilon, \quad (1)$$

where Y is a real-valued random variable defined on a probability space (Ω, \mathcal{B}, P) , $Z = (Z_1, \dots, Z_p)^T$ and U are observed explanatory variables and so-called index variables defined on the same probability space, respectively. $\beta(U) = (\beta_1(U), \dots, \beta_p(U))^T$ consists of p unknown varying-coefficient functions, $\{X(t) : t \in \mathcal{T}\}$ is a zero mean, the second-order stochastic process valued in $H = L^2(\mathcal{T})$, the set of all square-integrable functions on \mathcal{T} with the inner product $\langle x, y \rangle = \int_{\mathcal{T}} x(t)y(t)dt$, $\forall x, y \in H$ and norm $\|x\| = \langle x, x \rangle^{1/2}$, $\alpha(t) \in H$ is unknown slope function, the error term ε is assumed to be independent of (Z, U, X) with $E(\varepsilon|Z, U, X) = 0$. Without loss of generality, we assume that U takes values from the unit interval $[0, 1]$ and suppose throughout that $\mathcal{T} = [0, 1]$. The VCPFLRM is flexible and includes some existing statistical models. For example, when $\alpha(t) = 0$, it reduces to the varying coefficient model; when $\beta(\cdot)$ are all set as constant, it reduces to the partially FLM [15]; when $p \equiv 1$ and $Z_1 = 1$, it reduces to the semi-FLM [16].

To estimate the slope function $\alpha(\cdot)$ and the coefficient functions $\beta_j(\cdot)$, $j = 1, \dots, p$ in VCPFLRM, Peng et al. [14] proposed the least squares estimation procedure based on B-spline basis function approximations. Feng and Xue [17] proposed a profile least squares estimation by using functional principal component (FPC) analysis and local linear smoothing technique. However, all the above existing articles are built on least square method and the assumptions that the error is normally distributed, which are sensitive to outliers and heavy-tailed error distributions. Hence, robustness against outliers is a very important issue in the VCPFLRM. Recently, Wang et al. [18] investigated a new estimation method for linear regression models by minimizing exponential squared loss (ESL) $\rho(\cdot)$, i.e., $\rho(t) = 1 - \exp(-t^2/h)$, where the tuning parameter $h \in (0, \infty)$ controls the degree of robustness and efficiency for the estimators. For large h , $1 - \exp(-t^2/h) \approx t^2/h$, and therefore the proposed estimators are similar to the least squares estimators in the extreme case. For a small h , the outliers in the observations t will receive a small impact on the estimators. Therefore, this method is more robust to outliers than the least squares method. Due to its nice property, this method has been quickly extended to the statistical models for the sake of robust inference, the recent literature include but are not restricted to Jiang et al. [19] for the varying coefficient partially nonlinear model, Song et al. [20] for the single index varying coefficient model, Yu et al. [21] for the semi-functional linear regression model, Lv et al. [22] for the generalized varying coefficient partially linear model. Even though exponential-squared-based estimation has been well developed, the theory and methodology for VCPFLRM are none. It is therefore our impetus for applying this method to VCPFLRM.

The main objective of this article is to propose a new robust estimation procedure based on the ESL function for the VCPFLRM. To estimate $\alpha(\cdot)$ in VCPFLRM, two commonly used approaches have been proposed, including the FPC basis method and the basis function expansions (such as B-spline functions approximation) method. Since FPC has some desirable properties in approximating a smooth function, for example, it is a data-driven technique where the data can be represented as linear combinations of functions estimated from the data. Besides, by requiring the order of generalized Fourier coefficients based on the FPC basis, we can control the smoothness of the slope function. Due to these advantages mentioned above, we first use B splines and FPC basis to approximate the varying coefficient functions and slope function in model (1), respectively. With the help of these approximations, the estimation problem based on the ESL function for VCPFLRM becomes a problem of estimating the coefficients in the linear combinations. Under some mild regularity conditions, we show that the estimators of coefficient functions and slope functions can achieve the optimal convergence rate. Moreover, our numerical studies and a real data

application in the sequel demonstrate that our proposed method is at least comparable to some existing works. It is worth pointing out that there exist some essential differences between our proposed method and that of Wang et al. [18] and Yu et al. [21]. First, our studied VCPFLRM is flexible enough to cover the classical linear regression model and the semi-functional linear regression model as special cases. Accordingly, a different approach is required to deal with varying coefficient functions $\beta(\cdot)$, this difference leads to different theoretical properties for proposed estimators. Second, compared with the works of [18] and [21], though the estimate of $\beta(\cdot)$ and $\alpha(\cdot)$ can be easily obtained by transforming the original VCPFLRM to a standard linear regression model, we need to choose extra tuning parameters by the data-driven procedure for spline estimates of $\beta(\cdot)$, such as the number of interior knots, which controls the dimensions of the spline spaces used to approximate the true varying coefficient functions.

The rest of the article is organized as follows. In Section 2, we describe the proposed estimation procedure for VCPFLRM. In Section 3, we establish the theoretical properties of the resulted estimators. In Section 4, we develop an estimation algorithm and discuss the selection of tuning parameters. Section 5 illustrates the numerical performance of the proposed method through simulation experiments. A real data analysis is presented in Section 6, and some concluding remarks are followed in Section 7. All proofs of theorems are provided in the Appendix.

2 Estimation procedure

Let $\{Y_i, Z_i, U_i, X_{it}\}_{i=1}^n$ be an independent and identically distributed sample that is generated from model (1). Motivated by [18], we estimate the slope function $\alpha(t)$ and nonparametric coefficients $\beta(U)$ in VCPFLRMs by maximizing the following ESL function

$$\sum_{i=1}^n \exp \left\{ - \frac{(Y_i - Z_i^T \beta(U_i) - \int_0^1 \alpha(t) X_i(t) dt)^2}{h} \right\}. \quad (2)$$

In what follows, we first approximate the nonparametric components $\beta(U)$ by B splines. Let $B(u) = (B_1(u), \dots, B_K(u))^T$ be the B-spline basis functions, where $K = k_n + q + 1$ is the number of basis functions, q is the degree of the spline, and k_n is the number of internal knots that control the smoothness of nonparametric functions. Thus, for any function $\beta_j(u)$, $j = 1, \dots, p$, we have

$$\beta_j(u) \approx \sum_{k=1}^K \eta_{jk} B_k(u). \quad (3)$$

Note that it is possible to specify different K for each component, but we assume that they are the same for simplicity. Next, we construct the covariance function and the empirical covariance function for the random process $X(\cdot)$ as follows:

$$K(s, t) = \text{Cov}(X(s), X(t)), \quad \hat{K}(s, t) = \frac{1}{n} \sum_{i=1}^n X_i(s) X_i(t). \quad (4)$$

Obviously, the covariance function $K(s, t)$ defines a linear operator that maps a function f to Kf given by $(Kf)(s) = \int K(t, s) f(t) dt$. We assume that the linear operator with kernel $K(s, t)$ is positive definite. Let $\lambda_1 > \lambda_2 > \dots > 0$ and $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_{n+1} = \dots = 0$ be the ordered eigenvalue sequences of the linear operators with kernels $K(s, t)$ and $\hat{K}(s, t)$. $\{\phi_j\}$ and $\{\hat{\phi}_j\}$ are an orthonormal sequence of eigenfunctions, and it is clear that the vectors ϕ_1, ϕ_2, \dots form an orthonormal basis in H . Then, consider the spectral decomposition of the covariance function $K(s, t)$ and $\hat{K}(s, t)$ as follows:

$$K(s, t) = \sum_{j=1}^{\infty} \lambda_j \phi_j(s) \phi_j(t), \quad \hat{K}(s, t) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\phi}_j(s) \hat{\phi}_j(t). \quad (5)$$

According to the Karhunen-Loève presentation, two L^2 -valued functions $X(t)$ and $\alpha(t)$ have the following expansions:

$$X(t) = \sum_{i=1}^{\infty} \xi_i \phi_i(t), \quad \alpha(t) = \sum_{j=1}^{\infty} \gamma_j \phi_j(t), \quad (6)$$

where $\xi_i = \int_0^1 X(t) \phi_i(t) dt$, $\gamma_j = \int_0^1 \alpha(t) \phi_j(t) dt$. It follows that ξ_i s are uncorrelated random variables with mean 0 and variance $E(\xi_i^2) = \lambda_i$. For more details, see [23].

Let $\Pi_i = (Z_{i1}B_1(U_i), \dots, Z_{iK}B_K(U_i), \dots, Z_{ip}B_p(U_i))^T$, $\eta = (\eta_1^T, \dots, \eta_p^T)^T$ with $\eta_j = (\eta_{j1}, \dots, \eta_{jK})^T$. Based on equations (3)–(6), model (1) can be rewritten as follows:

$$Y_i \approx \Pi_i^T \eta + \sum_{j=1}^m \gamma_j \langle \phi_j, X_i \rangle + \varepsilon_i, \quad i = 1, \dots, n, \quad (7)$$

where $m \leq n$ is the truncation level that trades off approximation error against variability and typically diverges with n . Replace ϕ_j by $\hat{\phi}_j$ for $j = 1, \dots, m$, model (7) can be rewritten as follows:

$$Y_i \approx \Pi_i^T \eta + V_i^T \gamma + \varepsilon_i, \quad i = 1, \dots, n, \quad (8)$$

where $V_i = \{\langle X_i, \hat{\phi}_j \rangle\}_{j=1, \dots, m}$, $\gamma = (\gamma_1, \dots, \gamma_m)^T$. Thus, the objective function (2) becomes

$$Q_n(\eta, \gamma) = \sum_{i=1}^n \exp \left\{ -\frac{(Y_i - \Pi_i^T \eta - V_i^T \gamma)^2}{h} \right\}. \quad (9)$$

Denote the maximizer of (9) by $\hat{\eta}$ and $\hat{\gamma}$. Consequently, the estimated component functions are given by $\hat{\beta}_j(u) = B^T(u) \hat{\eta}_j$, $\hat{\alpha}(t) = \sum_{j=1}^m \hat{\gamma}_j \hat{\phi}_j(t)$.

3 Theoretical results

In this section, we aim to establish the theoretical properties of the resulting estimators. First, we need to make some definitions and assumptions. We assume that the true functions in model (1) are $\alpha_0(\cdot)$ and $\beta_{0j}(\cdot)$, $j = 1, \dots, p$. We use C to denote a generic positive constant that can vary from line to line. $a_n \sim b_n$ means that a_n/b_n is bounded away from zero and infinity as $n \rightarrow \infty$. The notation $\|\cdot\|$ is the L^2 norm for a function or the Euclidean norm for a vector. Let $S = \langle X(t), \alpha_0(t) \rangle$, $\psi_h(\varsigma) = \exp(-\varsigma^2/h)$, $F(z, u, s, h) = E(\psi_h''(\varepsilon)|Z = z, U = u, S = s)$, and $G(z, u, s, h) = E(\psi_h'(\varepsilon)^2|Z = z, U = u, S = s)$. The following regularity conditions will be required, which are common and some can be found in [14], [24].

(C1) $E\|X(\cdot)\|^4 < C < \infty$, $E[\xi_i^4] \leq C\lambda_i^2$, for $i \geq 1$, for $\iota > 2$, $E(|Z_j|^\iota) < \infty$ for $j = 1, \dots, p$.

(C2) The eigenvalues λ_i and score coefficients γ_j satisfy the following conditions, respectively.

(a) There exist some constant C and $a > 1$ such that

$$C^{-1}i^{-a} \leq \lambda_i \leq Ci^{-a}, \quad \lambda_i - \lambda_{i+1} \geq Ci^{-a-1}, \quad i \geq 1.$$

(b) There exist some constant C and $b > a/2 + 1$ such that

$$|\gamma_j| \leq Cj^{-b}, \quad j \geq 1.$$

(C3) The tuning parameter m satisfies $m \sim n^{1/(a+2b)}$.

(C4) The density function of U , $f_U(u)$ is bounded away from 0 and infinity on $[a, b]$. Furthermore, we assume that $f_U(u)$ is continuously differentiable on (a, b) .

(C5) $\beta_j(t) \in \mathcal{F}_r$, $j = 1, \dots, p$ for some $r > 2$, where \mathcal{F}_r is the collection of all functions on $[0, 1]$ whose d th order derivative satisfies the Hölder condition of the order ς with $r \equiv d + \varsigma$ and $0 < \varsigma \leq 1$.

(C6) $F(z, u, s, h)$ and $G(z, u, s, h)$ are continuous with respect to (z, u, s) and $F(z, u, s, h) < 0$ for any $h > 0$.

(C7) $E(\psi'_h(\varepsilon)|z, u, s) = 0$, $E(\psi''_h(\varepsilon)|z, u, s)$, $E(\psi'''_h(\varepsilon)|z, u, s)$, $E(\psi''''_h(\varepsilon)|z, u, s)$ are continuous with respect to (z, u, s) .

Theorem 1. Suppose that the regularity conditions (C1)–(C7) hold, and the number of interior knots $k_n \sim n^{\frac{1}{2r+1}}$, then we have

$$\|\hat{\alpha}(\cdot) - \alpha_0(\cdot)\|^2 = O_p(n^{\frac{-2r}{2r+1}} + n^{\frac{-2b-1}{a+2b}}), \quad \|\hat{\beta}_j(\cdot) - \beta_{0j}(\cdot)\|^2 = O_p(n^{\frac{-2r}{2r+1}} + n^{\frac{-2b-1}{a+2b}}), \quad j = 1, \dots, p.$$

Remark 1. Theorem 1 derives the convergence rates of the estimators $\hat{\alpha}(\cdot)$ and $\hat{\beta}_j(\cdot)$. When we take $m \sim k_n \sim n^{\frac{1}{2r+1}}$ and $\frac{1}{2} + a < r \leq \frac{a+2b-1}{2}$, then $\|\hat{\beta}_j(\cdot) - \beta_{0j}(\cdot)\|^2 = O_p(n^{\frac{-2r}{2r+1}})$ and $\|\hat{\alpha}(\cdot) - \alpha_0(\cdot)\|^2 = O_p(n^{\frac{-2r-a}{2r+1}})$, we see that $\hat{\beta}_j(\cdot)$ obtains the optimal global convergence rate, but $\hat{\alpha}(\cdot)$ obtains $O_p(n^{\frac{-2r-a}{2r+1}})$, the introduction of the varying coefficient component affects the convergence rate of the slope function $\alpha(\cdot)$. When we take $m \sim k_n \sim n^{\frac{1}{a+2b}}$ and $r > \frac{a+2b-1}{2}$, we show $\|\hat{\beta}_j(\cdot) - \beta_{0j}(\cdot)\|^2 = O_p(n^{\frac{-a+2b-1}{a+2b}})$ and $\|\hat{\alpha}(\cdot) - \alpha_0(\cdot)\|^2 = O_p(n^{\frac{-2b-1}{a+2b}})$, so $\hat{\alpha}(\cdot)$ attains the same convergence rate as those of the estimators of [10] and [15], which are optimal in the minimax sense. However, $\hat{\beta}_j(\cdot)$ do not obtain the optimal global convergence rate, so the convergence rates of varying coefficient components are also affected by $\alpha(\cdot)$. When we take $k_n \sim n^{\frac{1}{2r+1}} \sim m \sim n^{\frac{1}{a+2b}}$, and $r = \frac{2b-1}{2(a+1)}$, both the estimators $\hat{\alpha}(\cdot)$ and $\hat{\beta}_j(\cdot)$ obtain the optimal global convergence rates.

4 Computational issue

4.1 Algorithm implementation

In this subsection, we apply the Newton-Raphson iteration algorithm to solve the estimating equation (9) since the ESL is smooth with continuous second-order derivative. To be more specific, let $\Psi_i = (\Pi_i^T, V_i^T)^T$, $\theta = (\eta^T, \gamma^T)^T$, and $Q_n(\theta) = -\sum_{i=1}^n \exp\{-(Y_i - \Psi_i^T \theta)^2 / h\}$, then we outline the iterative algorithm as follows:

Step 1. Start with an initial estimator $\hat{\theta}^{(0)}$.

Step 2. Update the estimator $\hat{\theta}^{(k)}$ of θ by using the iterative procedure

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \left[\frac{\partial^2 Q_n(\theta)}{\partial \theta \partial \theta^T} \Big|_{\theta = \hat{\theta}^{(k)}} \right]^{-1} \times \left[\frac{\partial Q_n(\theta)}{\partial \theta} \Big|_{\theta = \hat{\theta}^{(k)}} \right].$$

Step 3. Iterate Step 2 until convergence and denote the final estimator of θ as $\hat{\theta}$.

4.2 Selection of tuning parameter h

In this subsection, we will address how to select the tuning parameter h for the proposed estimators in practice. For simplicity, we further assume that ε is independent of (Z, U, X) . Following the advice of [19], we can estimate $F(h)$ and $G(h)$ in implementation by

$$\hat{G}(h) = \frac{1}{n} \sum_{i=1}^n \frac{4}{h^2} \{\hat{\varepsilon}_i \exp(-\hat{\varepsilon}_i^2 / h)\}^2, \quad \hat{F}(h) = \frac{1}{n} \sum_{i=1}^n \frac{4}{h^2} \exp(-\hat{\varepsilon}_i^2 / h) (\hat{\varepsilon}_i^2 - h/2),$$

where $\hat{\varepsilon}_i = Y_i - \Pi_i^T \hat{\eta} - V_i^T \hat{\gamma}$, $\hat{\eta}$, and $\hat{\gamma}$ are estimated based on the pilot estimates. In practice, one can use some existing methods such as mean regression and median regression methods to obtain these pilot estimates. Therefore, the optimal h can be chosen as

$$\hat{h}_{\text{opt}} = \arg \min_h \{\hat{G}(h) \hat{F}^{-2}(h) \hat{\sigma}^2\},$$

where $\hat{\sigma}$ is estimated based on the pilot estimator. Then, we can use the grid search method recommended in [25] to find h , we find that the possible grid points for h can be selected from $0.5\hat{\sigma} \times 1.02^j$, $j = 1, 2, \dots, k$ for some fixed k , such as $k = 50$ or $k = 100$. In our numerical analysis, we set $k = 100$.

4.3 Extra parameter selections

To achieve good numerical performance, one also needs to choose the number of interior knots k_n and the tuning parameter m , appropriately. There are many methods such as the generalized cross-validation selector [26] and the BIC selector [27] can be used. In this work, we use cubic splines in all our numerical examples. Then use a BIC-type criterion to select k_n and m , that is,

$$\text{BIC}(k_n, m) = \log \left(\sum_{i=1}^n \left[1 - \exp \left\{ -\frac{(Y_i - \Pi_i^T \hat{\eta} - V_i^T \hat{\gamma})^2}{\hat{h}_{\text{opt}}} \right\} \right] \right) + \frac{\log n}{n} (p(k_n + 4) + m),$$

where $\hat{\eta}$ and $\hat{\gamma}$ are the proposed estimators by minimizing (9) with the number of interior knots k_n and the first m principal components. In this work, we select the tuning parameter and the number of interior knots by minimizing the above BIC criterion on the ranges $1 \leq m \leq 8$ and $[n^{1/9}] \leq k_n \leq [8n^{1/9}]$, respectively, where $[\varpi]$ denotes the largest integer not greater than ϖ .

5 Simulation experiments

In this section, we conduct simulations to compare the finite sample performance of our proposed estimation method (ESL) with the other two methods: the quantile regression method [24] with quantile $\tau = 0.5$ (QR) and the least square (LS) method [14]. We generated data from the model as follows:

$$Y = Z_1 \beta_1(U) + Z_2 \beta_2(U) + Z_3 \beta_3(U) + \int_{\tau} \alpha(t) X(t) dt + \varepsilon,$$

where the coefficient functions are $\beta_1(u) = 2 \sin(2\pi u)$, $\beta_2(u) = 8u(1 - u)$, and $\beta_3(u) = 5.5 + 10 \exp(2u - 1)$, the covariate $Z = (Z_1, Z_2, Z_3)^T$, with each Z_i , $i = 1, 2, 3$, being standard normal distribution and the correlation between Z_i and Z_j being $0.5^{|i-j|}$, $U \sim U(0, 1)$. For the functional linear component, we take the same form as [21], that is, $\alpha(t) = \sqrt{2} \sin(\pi t/2) + 3\sqrt{2} \sin(3\pi t/2)$ and $X(t) = \sum_{j=1}^{50} \xi_j v_j(t)$, where the ξ_j are independently distributed as the normal with mean 0 and variance $\lambda_j = 10((j - 0.5)\pi)^{-2}$, $v_j(t) = \sqrt{2} \sin((j - 0.5)\pi t)$. In order to examine the robustness and efficiency of our proposed method, the following three different error distributions are considered:

Case 1. ε follows the standard normal distribution, abbreviated as $N(0, 1)$.

Case 2. ε follows the standard t distribution with three degrees of freedom, abbreviated as $t(3)$, which is used to produce heavy-tailed distribution.

Case 3. ε follows the contaminated normal distribution with $k\%$ data from $N(0, \ell)$ and remaining $1 - k\%$ data from $N(0, 1)$, abbreviated as $0.9N(0, 1) + 0.1N(0, 10)$. Obviously, the $k\%$ data can be regarded as outliers, and we set $k = 10$ and $\ell = 10$ in this case.

To evaluate the performance of our estimation for slope function $\alpha(\cdot)$ and coefficient functions $\beta_j(\cdot)$, we define the following square root of the average square errors (RASE):

$$\begin{aligned} \text{RASE}_0 &= \sqrt{\frac{1}{n_{1, \text{grid}}} \sum_{i=1}^{n_{1, \text{grid}}} (\hat{\alpha}(t_i) - \alpha(t_i))^2}, \\ \text{RASE}_j &= \sqrt{\frac{1}{n_{2, \text{grid}}} \sum_{i=1}^{n_{2, \text{grid}}} (\hat{\beta}_j(u_i) - \beta_j(u_i))^2}, \quad j = 1, 2, 3, \end{aligned}$$

where $\{t_i, i = 1, \dots, n_{1, \text{grid}}\}$ and $\{u_i, i = 1, \dots, n_{2, \text{grid}}\}$ are the grid points at which the functions $\hat{\alpha}(\cdot)$ and $\hat{\beta}_j(\cdot)$ are evaluated, respectively. In our simulation experiments, we simply set $n_{1, \text{grid}} = 201$ and $n_{2, \text{grid}}$ equals to the sample size n in each scenario. A total of 200 replications with $n = 100, 200, 600$ are performed. The mean, median, and standard error (SE) of the RASE values for three error distribution cases are reported in Tables 1–3. Several observations can be found in Tables 1–3. First, when the error follows a standard normal distribution, the LS estimate slightly outperforms the QR and ESL estimates, and ESL performs slightly better than QR. Second, ESL is significantly superior to LS in the other two cases, and this is expected because the performance of LS is sensitive to heavy-tailed distributions or outliers, ESL seems to be slightly better than QR in most cases. Third, for the given error distribution, all the considered estimators become better and better as the sample size n increase gradually. Figure 1 demonstrates the performances of average curve estimates of $\beta_j(\cdot)$ and $\alpha(\cdot)$ when $t \sim 0.9N(0, 1) + 0.1N(0, 10)$ and $n = 100$. The plots for other cases are similar and are omitted. All these results indicate that our proposed method works well under the considered settings. To further demonstrate the performance of three different methods, we also consider the Cauchy error distribution. The results are given in Figure 2. We find LS has poor performance since the LS estimates will break down and suffer from consistency when the error is Cauchy.

To further compare the efficiency of the LS, QR, and ESL estimation of the nonparametric functions $\hat{g} = (\hat{\alpha}(\cdot), \hat{\beta}_j(\cdot))$, we consider the ratio of average squared error (RT) defined as

$$\text{ASE} = \left\{ \frac{1}{n_{1, \text{grid}}} \sum_{i=1}^{n_{1, \text{grid}}} (\hat{\alpha}(t_i) - \alpha(t_i))^2 + \frac{1}{n_{2, \text{grid}}} \sum_{j=1}^3 \sum_{i=1}^{n_{2, \text{grid}}} (\hat{\beta}_j(u_i) - \beta_j(u_i))^2 \right\}, \quad \text{RT}(\hat{g}) = \frac{\text{ASE}(\hat{g}_{\text{LS}})}{\text{ASE}(\hat{g})},$$

Table 1: Simulation results for comparison under $N(0, 1)$ error distribution when the sample size $n = 100, 200, 600$

n	Method		RASE ₀	RASE ₁	RASE ₂	RASE ₃
100	LS	Mean	0.4443	0.2916	0.2635	0.2743
		Median	0.4100	0.2823	0.2596	0.2628
		SE	0.1733	0.0844	0.0902	0.0948
	QR	Mean	0.4477	0.3480	0.3194	0.3263
		Median	0.4205	0.3261	0.3155	0.3227
		SE	0.1619	0.1036	0.1126	0.1154
	ESL	Mean	0.4354	0.3274	0.2952	0.3012
		Median	0.4105	0.3146	0.2868	0.2878
		SE	0.1472	0.1028	0.1124	0.1061
200	LS	Mean	0.2952	0.2137	0.1679	0.1665
		Median	0.2810	0.2081	0.1590	0.1614
		SE	0.0977	0.0491	0.0600	0.0582
	QR	Mean	0.3285	0.2473	0.2094	0.2004
		Median	0.3022	0.2427	0.1992	0.1901
		SE	0.1133	0.0628	0.0752	0.0673
	ESL	Mean	0.3146	0.2328	0.1865	0.1845
		Median	0.2972	0.2294	0.1773	0.1777
		SE	0.1033	0.0539	0.0711	0.0620
600	LS	Mean	0.1728	0.1529	0.0928	0.0921
		Median	0.1639	0.1505	0.0917	0.0887
		SE	0.0631	0.0266	0.0305	0.0299
	QR	Mean	0.1805	0.1785	0.1127	0.1126
		Median	0.1682	0.1748	0.1101	0.1113
		SE	0.0668	0.0284	0.0360	0.0371
	ESL	Mean	0.1762	0.1696	0.0993	0.1003
		Median	0.1670	0.1664	0.0984	0.0989
		SE	0.0624	0.0196	0.0321	0.0320

Table 2: Simulation results for comparison under $t(3)$ error distribution when the sample size $n = 100, 200, 600$

n	Method		$RASE_0$	$RASE_1$	$RASE_2$	$RASE_3$
100	LS	Mean	0.5620	0.4460	0.4235	0.4098
		Median	0.4624	0.4126	0.3908	0.3735
		SE	0.3453	0.2233	0.2226	0.1807
	QR	Mean	0.4670	0.4122	0.3796	0.3815
		Median	0.4487	0.3772	0.3554	0.3543
		SE	0.1551	0.1505	0.1506	0.1388
	ESL	Mean	0.4696	0.3964	0.3752	0.3777
		Median	0.4403	0.3666	0.3473	0.3443
		SE	0.1628	0.1452	0.1627	0.1637
200	LS	Mean	0.3350	0.2967	0.2652	0.2621
		Median	0.3080	0.2810	0.2608	0.2452
		SE	0.1280	0.0896	0.0951	0.1034
	QR	Mean	0.3101	0.2688	0.2351	0.2395
		Median	0.2884	0.2634	0.2256	0.2250
		SE	0.1006	0.0654	0.0798	0.0913
	ESL	Mean	0.3151	0.2615	0.2267	0.2238
		Median	0.2882	0.2522	0.2249	0.2114
		SE	0.1044	0.0721	0.0767	0.0852
600	LS	Mean	0.1944	0.2007	0.1579	0.1506
		Median	0.1783	0.1929	0.1523	0.1430
		SE	0.0887	0.0628	0.0572	0.0534
	QR	Mean	0.1736	0.1852	0.1287	0.1269
		Median	0.1582	0.1807	0.1259	0.1256
		SE	0.0631	0.0280	0.0420	0.0451
	ESL	Mean	0.1729	0.1747	0.1173	0.1150
		Median	0.1616	0.1714	0.1106	0.1130
		SE	0.0618	0.0247	0.0387	0.0381

where \hat{g}_{LS} denotes the LS estimate of g , and \hat{g} could be taken as the QR or ESL estimate. The corresponding results are summarized in Table 4.

From Table 4, we can see that the RTs of the ESL method are slightly smaller than 1 compared with the RTs of the QR method when the error is normally distributed, for the other two error distributions, the RTs of the ESL method are greater than 1, which indicates that our proposed estimation method can significantly improve the estimation efficiency.

6 A real data analysis

For illustration purposes, we apply the proposed methodology to analyze the spectral data. The data set contains 215 chopped meat samples, each data sample contains fat, protein, moisture contents, and spectral curve. The three contents are measured in percent and determined by analytic chemistry. The spectral curve consists of 100 wavelengths of absorbance spectrum records. Our main interest is to explore the relationship between fat content and other factors. In the subsequent analysis, we treat the fat content as the response variable Y , the protein content as Z , the moisture content as U , and the spectral curve as $X(t)$. Beforehand, we standardize the response variable Y and two other contents Z and U . As done in [17], we use the following VCPFLRM to fit the data:

Table 3: Simulation results for comparison under $0.9N(0, 1) + 0.1N(0, 10)$ error distribution when the sample size $n = 100, 200, 600$

n	Method		$RASE_0$	$RASE_1$	$RASE_2$	$RASE_3$
100	LS	Mean	0.8708	0.8677	0.7990	0.8095
		Median	0.6436	0.7485	0.6906	0.7061
		SE	0.7456	0.5025	0.4587	0.4369
	QR	Mean	0.4723	0.4329	0.4056	0.4057
		Median	0.4469	0.3674	0.3851	0.3685
		SE	0.1562	0.2029	0.1706	0.1859
	ESL	Mean	0.4591	0.3460	0.3182	0.3211
		Median	0.4318	0.3023	0.2916	0.2973
		SE	0.1732	0.1951	0.1416	0.1487
200	LS	Mean	0.4754	0.5300	0.5423	0.5536
		Median	0.4003	0.5009	0.5111	0.5247
		SE	0.2881	0.1910	0.2462	0.2259
	QR	Mean	0.3145	0.2692	0.2295	0.2424
		Median	0.2941	0.2645	0.2217	0.2360
		SE	0.1022	0.0647	0.0800	0.0795
	ESL	Mean	0.3073	0.2268	0.1875	0.1960
		Median	0.2836	0.2180	0.1799	0.1903
		SE	0.1086	0.0520	0.0637	0.0695
600	LS	Mean	0.2555	0.3250	0.2904	0.2970
		Median	0.2344	0.3193	0.2843	0.2930
		SE	0.1128	0.0859	0.1081	0.1156
	QR	Mean	0.1744	0.1898	0.1267	0.1278
		Median	0.1677	0.1858	0.1238	0.1244
		SE	0.0584	0.0286	0.0406	0.0415
	ESL	Mean	0.1700	0.1638	0.1026	0.1053
		Median	0.1596	0.1608	0.1008	0.1025
		SE	0.0609	0.0271	0.0326	0.0351

$$Y = \beta_0(U) + Z\beta_1(U) + \int_{850}^{1,050} \alpha(t)X(t)dt + \varepsilon. \quad (10)$$

For a fair evaluation, the randomly selected 129 samples are used as training data to build the model, and the remaining 86 observations are left as the validation data for prediction purposes. The prediction performance is measured by the mean squared error proposed by [28], which is defined as $MSE = \frac{1}{86} \sum_{i \in \mathcal{G}} (Y_i - \hat{Y}_i)^2$, where \mathcal{G} is the validation data set. To verify the proposed robust estimation procedure, we re-analyzed this data set by including some outliers in the training response variable; in other words, the response Y of training data with ι high leverage outliers being replaced by $Y + \kappa$ (Case 1: $\kappa = 5$, $\iota = 10$, Case 2: $\kappa = 10$, $\iota = 10$ and Case 3: $\kappa = 5$, $\iota = 20$). Then, we apply the three different estimation procedures (LS, QR with quantile $\tau = 0.5$, and ESL) to analyze the data set by VCPFLRM stated as (10). The mean values of MSE of the proposed ESL procedure with those of QR and LS methods based on 200-time repetitions are shown to be 0.022, 0.023, and 0.706 for Case 1, 0.019, 0.032, and 2.378 for Case 2, 0.035, 0.042, and 1.323 for Case 3, respectively. Consequently, our proposed ESL method is better than the others in terms of performance of prediction when data set contains outliers. For comparison, we also use the functional linear regression model to fit the data. We only consider case 2, that is, the response Y of training data has 10 high leverage outliers, with 200-time repetitions, the mean values of MSE based on ESL, QR, and LS are shown to be 0.067, 0.102, and 3.187. This result reveals that our proposed model and estimation method can improve the accuracy of the prediction when data sets are subjected to high outliers in the response.

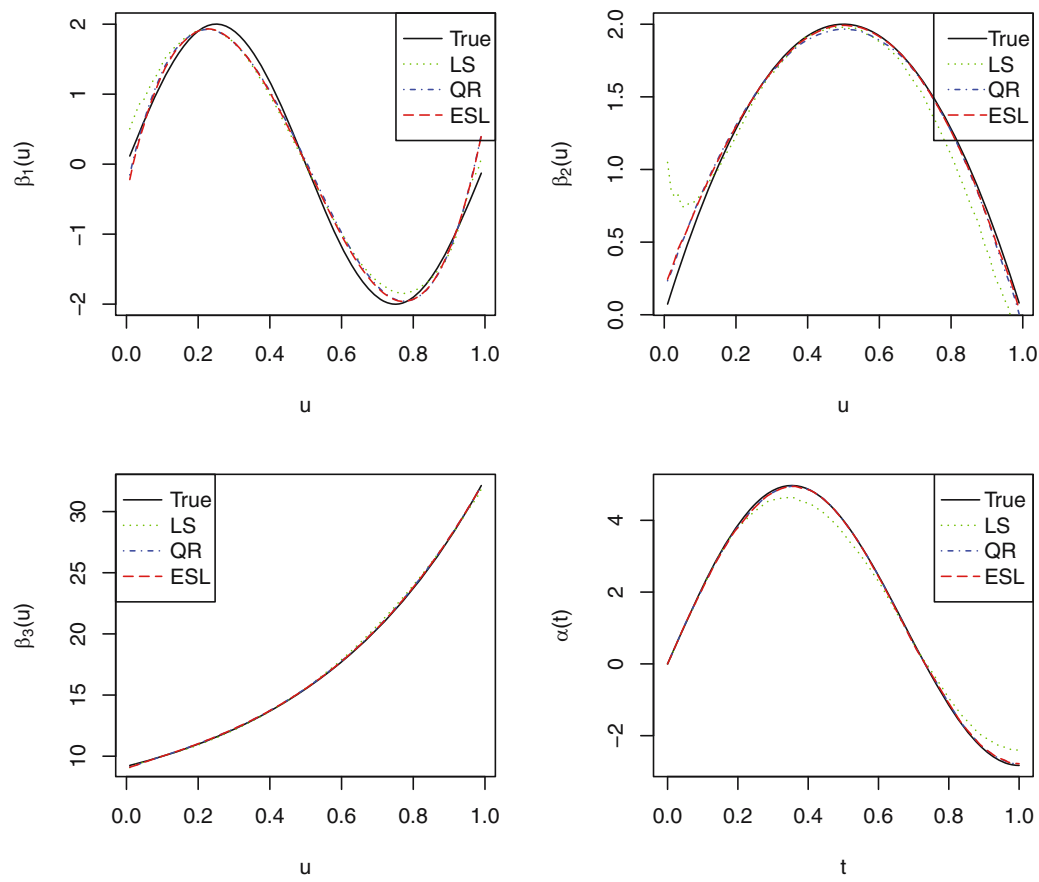


Figure 1: The estimated curves of functions $\beta_j(u)$, $j = 1, 2, 3$ and $\alpha(t)$ when $\varepsilon \sim 0.9N(0, 1) + 0.1N(0, 10)$ and $n = 100$.

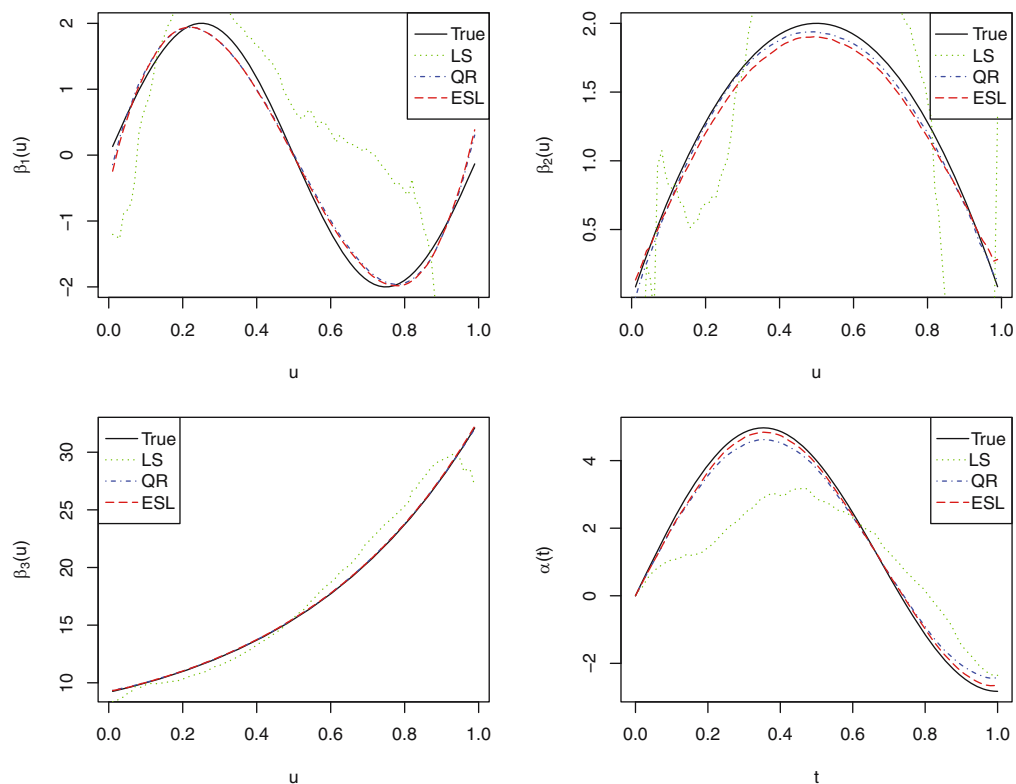


Figure 2: The estimated curves of functions $\beta_j(u)$, $j = 1, 2, 3$ and $\alpha(t)$ when $\varepsilon \sim \text{Cauchy}$ and $n = 100$.

Table 4: Mean values of the RT values

n	Method	$N(0, 1)$	$t(3)$	$0.9N(0, 1) + 0.1N(0, 10)$
100	LS	1	1	1
	QR	0.8919	1.1438	2.0084
	ESL	0.9515	1.1743	2.4272
200	LS	1	1	1
	QR	0.8646	1.1153	2.0176
	ESL	0.9245	1.1504	2.3424
600	LS	1	1	1
	QR	0.8821	1.1553	1.9079
	ESL	0.9435	1.2241	2.1859

7 Concluding remarks

In this article, we provide a robust and efficient estimation approach for VCPFLRMs based on the ESL function. The varying coefficients are approximated by B-splines, and the slope function is estimated by the FPC basis. The theoretical properties of proposed estimators are established under some conditions. Furthermore, we develop a computation algorithm to solve the optimization problems and then introduce a data-driven procedure to select the tuning parameters. The outstanding feature of the newly proposed method is that it can achieve good performance for a wide class of errors by selecting an appropriate tuning parameter h based on observed real data. The good properties of this method have also been demonstrated through both numeric examples and real data analysis. As discussed in [28], the FPC-based methods select data-driven basis functions relying on the decomposition of covariance function of the functional predictor $X(t)$. To compensate for this defect, we could use B-spline basis functions to approximate the slope function for its efficiency in function approximation and numerical computation. In addition, the method proposed here may be extended to other functional regression models, which could be considered in future research.

Acknowledgements: The authors are grateful to the editor, the associate editor, and the referees, whose comments and suggestions have greatly helped improve this article.

Conflict of interest: The authors state no conflict of interest.

References

- [1] J. Fan and W. Zhang, *Statistical methods with varying coefficient models*, Stat. Interface **1** (2008), 179–195, DOI: <https://dx.doi.org/10.4310/SII.2008.v1.n1.a15>.
- [2] L. Feng, C. Zou, Z. Wang, X. Wei, and B. Chen, *Robust spline-based variable selection in varying coefficient model*, Metrika **78** (2015), 85–118, DOI: <https://doi.org/10.1007/s00184-014-0491-y>.
- [3] C. Guo, H. Yang, and J. Lv, *Robust variable selection in high-dimensional varying coefficient models based on weighted composite quantile regression*, Statist. **58** (2017), 1009–1033, DOI: <https://doi.org/10.1007/s00362-015-0736-5>.
- [4] T. Hastie and R. Tibshirani, *Varying-coefficient models*, J. R. Stat. Soc. Ser. B. Methodol. **55** (1993), no. 4, 757–796, DOI: <https://doi.org/10.1111/j.2517-6161.1993.tb01939.x>.
- [5] J. Huang, C. Wu, and L. Zhou, *Varying-coefficient models and basis function approximations for the analysis of repeated measurements*, Biometrika **89** (2002), no. 1, 111–128, DOI: <https://doi.org/10.1093/biomet/89.1.111>.
- [6] H. Wang and Y. Xia, *Shrinkage estimation of the varying coefficient model*, J. Amer. Statist. Assoc. **104** (2009), no. 486, 747–757, DOI: <https://doi.org/10.1198/jasa.2009.0138>.
- [7] T. Cai and P. Hall, *Prediction in functional linear regression*, Ann. Statist. **34** (2006), no. 5, 2159–2179, DOI: <https://doi.org/10.1214/009053606000000830>.
- [8] T. Cai and M. Yuan, *Minimax and adaptive prediction for functional linear regression*, J. Amer. Statist. Assoc. **107** (2012), no. 499, 1201–1216, DOI: <https://doi.org/10.1080/01621459.2012.716337>.

- [9] H. Cardot, F. Ferraty, and P. Sarda, *Spline estimators for the functional linear model*, Statist. Sinica **13** (2003), no. 3, 571–591, <http://www.jstor.org/stable/24307112>.
- [10] P. Hall and J. Horowitz, *Methodology and convergence rates for functional linear regression*, Ann. Statist. **35** (2007), no. 1, 70–91, DOI: <https://doi.org/10.1214/009053606000000957>.
- [11] K. Kato, *Estimation in functional linear quantile regression*, Ann. Statist. **40** (2012), no. 6, 3108–3136, DOI: <https://doi.org/10.1214/12-AOS1066>.
- [12] J. Lei, *Adaptive global testing for functional linear models*, J. Amer. Statist. Assoc. **109** (2014), no. 506, 624–634, DOI: <https://doi.org/10.1080/01621459.2013.856794>.
- [13] F. Yao, H. Müller, and J. Wang, *Functional linear regression analysis for longitudinal data*, Ann. Statist. **33** (2005), no. 6, 2873–2903, DOI: <https://doi.org/10.1214/009053605000000660>.
- [14] Q. Peng, J. Zhou, and N. Tang, *Varying coefficient partially functional linear regression models*, Statist. Papers **57** (2016), 827–841, DOI: <https://doi.org/10.1007/s00362-015-0681-3>.
- [15] H. Shin, *Partial functional linear regression*, J. Statist. Plann. Inference **139** (2009), no. 10, 3405–3418, DOI: <https://doi.org/10.1016/j.jspi.2009.03.001>.
- [16] J. Zhou and M. Chen, *Spline estimators for semi-functional linear model*, Statist. Probab. Lett. **82** (2012), no. 3, 505–513, DOI: <https://doi.org/10.1016/j.spl.2011.11.027>.
- [17] S. Feng and L. Xue, *Partially functional linear varying coefficient model*, Statistics **50** (2016), no. 4, 717–732, DOI: <https://doi.org/10.1080/02331888.2016.1138954>.
- [18] X. Wang, Y. Jiang, M. Huang, and H. Zhang, *Robust variable selection with exponential squared loss*, J. Amer. Statist. Assoc. **108** (2013), no. 502, 632–643, DOI: <https://doi.org/10.1080/01621459.2013.766613>.
- [19] Y. Jiang, Q. Ji, and B. Xie, *Robust estimation for the varying coefficient partially nonlinear models*, J. Comput. Appl. Math. **326** (2017), 31–43, DOI: <https://doi.org/10.1016/j.cam.2017.04.028>.
- [20] Y. Song, L. Jian, and L. Lin, *Robust exponential squared loss-based variable selection for high-dimensional single-index varying-coefficient model*, J. Comput. Appl. Math. **308** (2016), 330–345, DOI: <https://doi.org/10.1016/j.cam.2016.05.030>.
- [21] P. Yu, Z. Zhu, and Z. Zhang, *Robust exponential squared loss-based estimation in semi-functional linear regression models*, Comput. Stat. **34** (2019), 503–525, DOI: <https://doi.org/10.1007/s00180-018-0810-2>.
- [22] J. Lv, H. Yang, and C. Guo, *Robust smooth-threshold estimating equations for generalized varying-coefficient partially linear models based on exponential score function*, J. Comput. Appl. Math. **280** (2015), 125–140, DOI: <https://doi.org/10.1016/j.cam.2014.11.003>.
- [23] L. Horváth and P. Kokoszka, *Inference for Functional Data with Applications*, Springer, New York, 2012.
- [24] P. Yu, J. Du, and Z. Zhang, *Varying-coefficient partially functional linear quantile regression models*, J. Korean Stat. Soc. **46** (2017), no. 3, 462–475, DOI: <https://doi.org/10.1016/j.jkss.2017.02.001>.
- [25] W. Yao, B. Lindsay, and R. Li, *Local modal regression*, J. Nonparametr. Stat. **24** (2012), no. 3, 647–663, DOI: <https://doi.org/10.1080/10485252.2012.678848>.
- [26] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Amer. Statist. Assoc. **96** (2001), no. 456, 1348–1360, DOI: <https://doi.org/10.1198/016214501753382273>.
- [27] H. Wang, R. Li, and C. Tsai, *Tuning parameter selectors for the smoothly clipped absolute deviation method*, Biometrika **94** (2007), no. 3, 553–568, DOI: <https://doi.org/10.1093/biomet/asm053>.
- [28] L. Huang, H. Wang, H. Cui, and S. Wang, *Sieve M-estimator for a semi-functional linear model*, Sci. China Math. **58** (2015), no. 11, 2421–2434, DOI: <https://doi.org/10.1007/s11425-015-5040-2>.
- [29] L. Schumaker, *Spline Functions: Basic Theory*, Wiley, New York, 1981.

Appendix

In the proofs, we use C to denote a generic constant that might assume different values at different places.

Proof of Theorem 1. Suppose $B^T(u)\eta_{0j}$ is the best approximating spline function for $\beta_j(u)$. Under the conditions (C4) and (C5) together with a constant C , we have

$$\sup_{u \in [0,1]} |\beta_j(u) - B^T(u)\eta_{0j}| \leq Ck_n^{-r}.$$

Let $\delta = n^{-r/(2r+1)} + n^{-(2b-1)/(2(a+2b))}$, $s_1 = \delta^{-1}(\eta - \eta_0)$, $s_2 = \delta^{-1}(\gamma - \gamma_0)$ with $\eta_0 = (\eta_{01}^T, \dots, \eta_{0p}^T)^T$ and $s = (s_1^T, s_2^T)^T$, and $\theta_0 = (\eta_0^T, \gamma_0^T)^T$. We want to show that for any given $\varrho > 0$, there exists a sufficiently large constant C such that

$$P\left\{\sup_{\|s\|=C} Q_n(\theta_0 + \delta s) < Q_n(\theta_0)\right\} \geq 1 - \varrho. \quad (\text{A1})$$

This implies that there is a local minimum in the ball $\{\theta_0 + \delta s : \|s\| \leq C\}$ with probability tending to 1. Using the Taylor expansion, we have

$$\begin{aligned} Q_n(\theta_0 + \delta s) - Q_n(\theta_0) &\leq -\delta \sum_{i=1}^n \psi'_h\{\varepsilon_i + Z_i^T R_{1i} + R_{2i}\} \Omega_i + \frac{1}{2} \delta^2 \sum_{i=1}^n \psi''_h\{\varepsilon_i + Z_i^T R_{1i} + R_{2i}\} \Omega_i^2 - \frac{1}{6} \delta^3 \sum_{i=1}^n \psi'''_h\{\zeta_i\} \Omega_i^3 \\ &\equiv I_1 + I_2 + I_3, \end{aligned}$$

where ζ_i is between $\varepsilon_i + Z_i^T R_{1i} + R_{2i}$ and $\varepsilon_i + Z_i^T R_{1i} + R_{2i} - \delta \Omega_i$, $\Omega_i = \Pi_i s_1 + V_i s_2$, $R_{1i} = \sum_{j=1}^p \beta_{0j}(U_i) - \Pi_i^T \eta_0$, $R_{2i} = \langle X_i(t), \alpha_0(t) \rangle - V_i^T \gamma_0$. From Conditions (C4) and (C5) and Corollary 6.21 in [29], we can derive that $R_{1i} = O_p(k_n^{-r})$.

Next, we consider R_{2i} . By $\|\phi_j - \hat{\phi}_j\|^2 = O_p(n^{-1j^2})$ (see, e.g., [15,21]), one has

$$|R_{2i}|^2 = \left| \int_0^1 X_i(t) \alpha_0(t) dt - V_i^T \gamma_0 \right|^2 \leq 2 \left| \sum_{j=1}^m \langle X_i, \hat{\phi}_j - \phi_j \rangle \gamma_{0j} \right|^2 + 2 \left| \sum_{j=m+1}^{\infty} \langle X_i, \phi_j \rangle \gamma_{0j} \right|^2 \triangleq 2A_1 + 2A_2.$$

For A_1 , by Conditions (C1) and (C2) and the Hölder inequality, it is obtained as follows:

$$A_1 = \left| \sum_{j=1}^m \langle X_i, \phi_j - \hat{\phi}_j \rangle \gamma_{0j} \right|^2 \leq cm \sum_{j=1}^m \|\phi_j - \hat{\phi}_j\|^2 |\gamma_{0j}|^2 \leq cm \sum_{j=1}^m O_p(n^{-1j^2-2b}) = O_p\left(n^{-\frac{a+4b-4}{a+2b}}\right) = o_p(n^{-(2b-1)(a+2b)}).$$

As for A_2 , due to

$$\begin{aligned} E\left\{\sum_{j=m+1}^{\infty} \langle X_i, \phi_j \rangle \gamma_{0j}\right\} &= 0, \\ \text{Var}\left\{\sum_{j=m+1}^{\infty} \langle X_i, \phi_j \rangle \gamma_{0j}\right\} &= \sum_{j=m+1}^{\infty} \lambda_j \gamma_{0j}^2 \leq c \sum_{j=m+1}^{\infty} j^{-(a+2b)} = O(n^{-\frac{a+2b-1}{a+2b}}), \end{aligned}$$

one has $A_2 = O_p\left(n^{-\frac{a+2b-1}{a+2b}}\right)$. Taking these together, we have $|R_{2i}|^2 = O_p\left(n^{-\frac{2b-1}{a+2b}}\right)$.

Moreover, by Taylor expansion, we have

$$\sum_{i=1}^n \psi'_h\{\varepsilon_i + Z_i^T R_{1i} + R_{2i}\} \Omega_i = \sum_{i=1}^n \{\psi'_h(\varepsilon_i) + \psi''_h(\varepsilon_i)[Z_i^T R_{1i} + R_{2i}] + \psi'''_h(\varepsilon_i^*)[Z_i^T R_{1i} + R_{2i}]^2\} \Omega_i,$$

where ε_i^* is between ε_i and $Z_i^T R_{1i} + R_{2i}$. Invoking conditions (C1) and (C7) and after some direct calculations, we obtain $\|I_1\| = O_p(n\delta^2\|s\|)$. For I_2 , we can prove $I_2 = E\{F(Z, U, S, h)\} O_p(n\delta^2\|s\|^2)$. Therefore, choosing a sufficiently large C , I_2 dominates I_1 uniformly $\|s\| = C$. Similarly, we can prove that $I_3 = O_p(n\delta^3\|s\|^3)$. Since $k_n \sim n^{1/(2r+1)}$, $m \sim n^{1/(a+2b)}$, it follows that $\delta\|s\| \rightarrow 0$ with $\|s\| = C$, which leads to $I_3 = o_p(I_2)$. Therefore, I_3 is

also dominated by I_2 in $\|s\| = C$. Hence, by choosing a sufficiently large C together with the condition (C6), (A.1) holds, and there exist local maximizers $\hat{\eta}$ and $\hat{\gamma}$ such that $\|\hat{\eta} - \eta_0\| = O_p(\delta)$, $\|\hat{\gamma} - \gamma_0\| = O_p(\delta)$.

Let $\Xi = \int_0^1 B(u)B^T(u)du$ and $R_{ij}(u) = \beta_{0j}(u) - B^T(u)\eta_{0j}$, we have

$$\begin{aligned}\|\hat{\beta}_j(\cdot) - \beta_{0j}(\cdot)\|^2 &= \int_0^1 (\hat{\beta}_j(u) - \beta_{0j}(u))^2 du \\ &= \int_0^1 \{B_j^T(u)\hat{\eta}_j - B_j^T(u)\eta_{0j} + R_{ij}(u)\}^2 du \\ &\leq 2 \int_0^1 \{B_j^T(u)\hat{\eta}_j - B_j^T(u)\eta_{0j}\}^2 du + 2 \int_0^1 R_{ij}(u)^2 du \\ &= 2(\hat{\eta}_j - \eta_{0j})^T \Xi (\hat{\eta}_j - \eta_{0j}) + 2 \int_0^1 R_{ij}(u)^2 du.\end{aligned}$$

Then, invoking $\|\Xi\| = O(1)$, a simple calculation yields $(\hat{\eta}_j - \eta_{0j})^T \Xi (\hat{\eta}_j - \eta_{0j}) = O_p(\delta^2)$. In addition, it is easy to show $\int_0^1 R_{ij}(u)^2 du = O_p\left(n^{-\frac{2r}{2r+1}}\right)$. Thus, we have $\|\hat{\beta}_j(\cdot) - \beta_{0j}(\cdot)\|^2 = O_p(\delta^2)$.

Observe that

$$\begin{aligned}\|\hat{\alpha}(t) - \alpha_0(t)\|^2 &= \left\| \sum_{j=1}^m \hat{\gamma}_j \hat{\phi}_j - \sum_{j=1}^{\infty} \gamma_{0j} \phi_j \right\|^2 \\ &\leq 2 \left\| \sum_{j=1}^m \hat{\gamma}_j \hat{\phi}_j - \sum_{j=1}^m \gamma_{0j} \phi_j \right\|^2 + 2 \left\| \sum_{j=m+1}^{\infty} \gamma_{0j} \phi_j \right\|^2 \\ &\leq 4 \left\| \sum_{j=1}^m (\hat{\gamma}_j - \gamma_{0j}) \hat{\phi}_j \right\|^2 + 4 \left\| \sum_{j=1}^m \gamma_{0j} (\hat{\phi}_j - \phi_j) \right\|^2 + 2 \sum_{j=m+1}^{\infty} \gamma_{0j}^2 \triangleq 4J_1 + 4J_2 + 2J_3.\end{aligned}$$

By Condition (C2) and the orthogonality of $\{\hat{\phi}_j\}$ and $\|\phi_j - \hat{\phi}_j\|^2 = O_p(n^{-1j^2})$, one has

$$J_1 = \left\| \sum_{j=1}^m (\hat{\gamma}_j - \gamma_{0j}) \hat{\phi}_j \right\|^2 \leq \sum_{j=1}^m |\hat{\gamma}_j - \gamma_{0j}|^2 = \|\hat{\gamma} - \gamma_0\|^2 = O_p(\delta^2). \quad (\text{A2})$$

$$\begin{aligned}J_2 &= \left\| \sum_{j=1}^m \gamma_{0j} (\hat{\phi}_j - \phi_j) \right\|^2 \leq m \sum_{j=1}^m \|\hat{\phi}_j - \phi_j\|^2 \gamma_{0j}^2 \leq \frac{m}{n} O_p\left(\sum_{j=1}^m j^2 \gamma_{0j}^2\right) \\ &= O_p\left(n^{-1} m \sum_{j=1}^m j^{2-2b}\right) = O_p(n^{-1} m) = o_p\left(n^{-\frac{2b-1}{a+2b}}\right),\end{aligned} \quad (\text{A3})$$

and

$$J_3 = \sum_{j=m+1}^{\infty} \gamma_{0j}^2 \leq C \sum_{j=m+1}^{\infty} j^{-2b} = O(n^{-\frac{2b-1}{a+2b}}). \quad (\text{A4})$$

Then, combining equations (A.2)–(A.4), we have $\|\hat{\alpha}(t) - \alpha(t)\|^2 = O_p(\delta^2) = O_p\left(n^{-\frac{2r}{2r+1}} + n^{-\frac{2b-1}{a+2b}}\right)$. Thus, we complete the proof of Theorem 1. \square