Statistica Sinica Preprint No: SS-2018-0483	
Title	Robust Inference in Varying-coefficient Additive Models for Longitudinal/Functional Data
Manuscript ID	SS-2018-0483
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0483
Complete List of Authors	Lixia Hu Tao Huang and Jinhong You
Corresponding Author	Tao Huang
E-mail	huang.tao@mail.shufe.edu.cn

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

Statistica Sinica

ROBUST INFERENCE IN VARYING-COEFFICIENT ADDITIVE MODELS FOR LONGITUDINAL/FUNCTIONAL DATA

Lixia Hu[1], Tao Huang[2] and Jinhong You[2]

[1] *Shanghai Lixin University of Accounting and Finance*

[2] *Shanghai University of Finance and Economics*

Abstract: This study provides a robust inference for a varying-coefficient additive model for sparse or dense longitudinal/functional data. A spline-based three-step M-estimation method is proposed for estimating the varying-coefficient component functions and the additive component functions. In addition, the consistency and asymptotic normality of sparse data and dense data are investigated within a unified framework. Furthermore, employing a regularized M-estimation method, a model identification procedure is proposed that consistently identifies an additive term and a varying-coefficient term. Simulation studies are used to evaluate the finite-sample performance of the proposed methods, and confirm the asymptotic theory. Lastly, real-data examples demonstrate the applicability of the proposed

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

methods.

Key words and phrases: B-spline, M-estimator, SCAD, tensor product, varying-coefficient additive model.

1. Introduction

Repeated-measurement data arise often in clinical, biometrical, epidemiological, social, and economic research (Diggle, Liang, and Zeger , 1994). Here, longitudinal and functional data are particularly common, and have different sampling mechanisms. Typically, data are termed functional when they are recorded densely over time in a continuum without noise, and are termed longitudinal when the measurements are made at a few discrete time points and include experimental error. However, in practice, functional data are analyzed after smoothing noisy observations (Ramsay and Ramsey , 2002). A vast body of literature considers statistical inferences for functional data that are based on observations at discrete time points and are contaminated with measurement errors, a practice that makes it possible to analyze longitudinal data and functional data within a unified framework (Li and Hsing, 2010; Yao , 2007). Others have studied longitudinal data using a functional principal components analysis (Yao, Müller and Wang , 2005).

In a typical repeated-measurement-data setting, a sample of n subjects or curves is observed at n_i discrete time points. If each n_i exceeds some power

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

of n, then the data are referred to as dense data. If each n_i is bounded by a finite positive number or follows a fixed distribution, then the data are referred to as sparse data. Recently, Zhang and Wang (2016) considered nonparametric estimations of the mean and covariance functions for sparse and dense functional data within a unified framework, where they categorized the data as sparse, dense, or ultra-dense, based on the magnitude of n_i relative to n.

Many studies have investigated nonparametric regression methods for functional data and longitudinal data with sparsity or/and denseness. Because of their simplicity, flexibility, and interpretability, varying-coefficient models (VCMs) have been used extensively to analyze longitudinal data (Hoover et al. , 1998; Xue and Zhu , 2007). Additive models (AMs) provide an alternative regression method (Carroll et al. , 2009; Xue, Qu, and Zhou , 2010). Here, Zhang, Park, and Wang (2013) proposed a time-varying AM for analyzing longitudinal data to capture dynamic effects. Recently, for analyzing functional data, Zhang and Wang (2015) proposed a novel nonparametric regression method called the varying-coefficient additive model (VCAM), which includes the classical AMs and VCMs as special cases. Specifically, let $Y(t)$ be a smooth random response process and $\mathbf{X} = (X_1, ..., X_p)^\tau$ be a p-vector of covariates. The regression function $m(t, \mathbf{x}) := \mathrm{E}[Y(t)|\mathbf{X} = \mathbf{x}]$ of a VCAM has the form

$$m(t, \mathbf{x}) = \alpha_0(t) + \sum_{k=1}^{p} \alpha_k(t)\beta_k(x_k), \tag{1.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

where α_k is the varying-coefficient component function, and β_k is the additive component function.

Zhang and Wang (2015) proposed a two-step spline estimation method for varying-coefficient component functions and additive component functions, based upon two key assumptions: (i) each subject (smooth process or function curve) is observed at dense time points; and (ii) each predictor is subject specific, but independent of the observation time. The above conditions are easily satisfied for functional data, but are restrictive for longitudinal data. Furthermore, if conditions (i) and/or (ii) are violated, then the estimation method of Zhang and Wang either fails or performs poorly, as shown in Table 6 of the Supplementary Material. In this study, we consider two real-data examples, namely, the CD4 cell count in HIV seroconversion (Zeger and Diggle , 1994), and the cigarette data set from the R package "phtt" (Bada and Liebl , 2012), which we investigate in the Supplementary Material. Note that each example violates condition (i) and/or (ii), meaning that the two-step spline estimation method proposed by Zhang and Wang (2015) is not appropriate. One of our aims herein is to relax conditions (i) and (ii), and to develop a general estimation method that has wider application in practical fields.

Although much of the literature focuses on the classical mean regression method, the method is sensitive to outliers and nonnormal error distributions.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

An alternative is the M-type robust regression method, which can treat mean, median, quantile, and more general robust-type regression methods within a unified framework. Many scholars have considered robust regression techniques, such as Koenker and Bassett (1978) for quantile regressions of linear models, He and Shi (1994) and He, Zhu, and Fung (2002) for M-estimators of partially linear models, and Tang and Cheng (2008) for M-estimators of VCMs.

Here, we consider a robust inference for a VCAM for sparse and dense longitudinal or functional data, allowing the predictors to be smooth processes covering condition (ii). We propose spline-based three-step M-estimators for varying-coefficient component functions and additive component functions. The asymptotic properties of the newly proposed estimators are presented within a unified framework, and we separate sparse data and dense data based on the relative order of n_i to n, which can be viewed as a generalization of Zhang and Wang (2016) to a VCAM. Similarly to Hu, Huang, and You (2018), a remarkable aspect of our estimators is the oracle property, which implies that the iteration step does not cause additional asymptotic errors. Furthermore, from the perspective of model parsimony, we develop a spline-based penalized M-estimator to decide whether the product term in (1.1) reduces to a varying-coefficient term or to an additive term, corresponding to an additive component function of linear form or a constant varying-coefficient component function, respectively. We also show

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

that an additive term and a varying-coefficient term can be selected correctly with probability approaching unity, under mild conditions.

The remainder of the paper is organized as follows. In Section 2, we describe the model setup and propose the spline-based three-step M-estimators for univariate component functions. In Section 3, we present the asymptotic theory for the proposed estimators. In Section 4, we introduce a robust model identification procedure, and in Section 5, we select the smoothing parameters. In Section 6, we use simulation examples to investigate the finite-sample performance, and use empirical examples to demonstrate the applicability of the proposed method. Finally, Section 7 concludes the paper. All technical proofs and additional numerical studies are relegated to the Supplementary Material.

2. Model and Estimation Method

2.1. Model assumptions

Let $Y(t)$ be a smooth response process and $\mathbf{X}(t) = \{X_1(t), ..., X_p(t)\}^\tau$ be a p-vector of smooth processes of covariates, where the superscript τ denotes the transpose of a vector or matrix. Without loss of generality, we assume that the response and covariates from a subject are L_2-integrable stochastic processes on the interval $[0, 1]$. The relationship between the response and the covariates is modeled by a VCAM, as follows:

$$Y(t) = \alpha_0(t) + \sum_{k=1}^{p} \alpha_k(t)\beta_k(X_k(t)) + U(t), \tag{2.1}$$

where $U(t)$ is the stochastic component of response process $Y(t)$, independent of covariate process $\mathbf{X}(t)$, with mean function $\mathrm{E}[U(t)] = 0$ and auto-covariance function $\gamma(t,s) = \mathrm{E}[U(t)U(s)]$. To uniquely identify univariate component functions, we impose the identification conditions $\int_0^1 \alpha_k(t)\mathrm{d}t = 1$ and $\mathrm{E}[\beta_k(X_k(t))] = 0$, following common practice for nonparametric regressions (Zhang and Wang , 2015; Wang and Yang , 2007; Vogt , 2012; Hu, Huang, and You , 2018).

In practical applications, the process Y is not observable, but can be measured at any given time with random error e, such that $\mathrm{E}(e) = 0$, $\mathrm{Var}(e) = \sigma_e^2$. We sample n subjects independently, and observe subject i at n_i time points $(t_{i1}, ..., t_{in_i})$, denoting y_{ij} and $\mathbf{x}_{ij} = (x_{ij1}, ..., x_{ijp})^\tau$ as the observations of the response and the vector of covariates at time t_{ij}, respectively. Then, the sample version of VCAM (2.1) can be written as

$$y_{ij} = \alpha_0(t_{ij}) + \sum_{k=1}^{p} \alpha_k(t_{ij})\beta_k(x_{ijk}) + U_{ij} + e_{ij}, \tag{2.2}$$

where $U_{ij} = U_i(t_{ij})$ is a realization of the subject-specific random trajectory $U_i(t)$ at observation time t_{ij}, and e_{ij} are independent and identical copies of the random measurement error e. As in Zhang and Wang (2015), we ignore the intra-subject covariance structure, and instead incorporate the covariance of $\{U_{ij}, j = 1, ..., n_i\}$ into the random error term, denoted as $\varepsilon_{ij} = U_{ij} + e_{ij}$.

Remark 1. The product term $\alpha_k(t)\beta_k(x_k)$ in VCAM (2.1) reduces to an additive term if α_k is a constant, and to a varying-coefficient term if β_k is a linear function.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

In other words, a VCAM is more flexible than either an AM or a VCM, and can greatly reduce the systematic bias of modeling.

2.2. Three-step M-estimation method

The spline method is a useful tool for fitting smooth nonparametric functions, and the B-spline basis is preferred for its computational stability. Let $\{\tilde{b}_1(x), ..., \tilde{b}_{K+m}(x)\}$ be a normalized m-order B-spline basis with K interior knots (De Boor , 1978). The scaled version of $\tilde{b}_k(x)$ is given by $b_k(x) = \sqrt{K+m}\tilde{b}_k(x)$, the favorable properties of which are presented in the Supplementary Material. Furthermore, similarly to Wang and Yang (2007), we construct a centralized version, represented as $\{B_1(x), ..., B_{K+m-1}(x)\}$. Under the assumption that both $\alpha_k(\cdot)$ and $\beta_k(\cdot)$ are $r(\leq q)$-order smooth, we adopt a q-order B-spline function to fit a univariate nonparametric function. For any $t \in [0,1]$ and x in the domain of $\beta_k(\cdot)$, we use the B-spline bases $\mathbf{b}_{\mathrm{C}}(t) = \{b_1(t), ..., b_{J_{\mathrm{C}}}(t)\}^\tau$ to approximate the varying-coefficient component function $\alpha_k(t)$; then, we use $\mathbf{B}_{k,\mathrm{A}}(x) = \{B_{k1}(x), ..., B_{kJ_{\mathrm{A}}}(x)\}^\tau$ to approximate the additive component function $\beta_k(x)$ for each $k = 1, .., p$, where J_{C} and J_{A} denote a sum of smooth degree r and the number of interior knots, respectively. The tensor product of $\mathbf{B}_{k,\mathrm{A}}(x_k)$ and $\mathbf{b}_{\mathrm{C}}(t)$ is defined as $\mathcal{T}_k(t, x_k) = \mathbf{B}_{k,\mathrm{A}}(x_k) \otimes \mathbf{b}_{\mathrm{C}}(t)$, where $\otimes$ represents the Kronecker product of matrices or vectors.

Now, we propose a spline-based three-step M-estimation method. Specifi-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

cally, we first obtain estimators for the varying-coefficient component functions. Then, we obtain an approximated AM and VCM by substituting the resultant estimators into VCAM (2.2). In this way, we estimate the varying-coefficient component functions and additive component functions.

Step I: Initial M-estimators of varying-coefficient component functions

In this step, we assume that B-spline bases have $\hbar_{\mathrm{C}}$ and $\hbar_{\mathrm{A}}$ interior knots for α_k and β_k, respectively. Using the tensor product of B-spline bases, the bivariate function $g_k(t,x_k)=\alpha_k(t)\beta_k(x_k)$ can be approximated as $g_k(t,x_k)\approx\gamma_k^\tau\mathcal{T}_k(t,x_k)$, where γ_k is a $\{(q+\hbar_{\mathrm{C}})(q+\hbar_{\mathrm{A}}-1)\}$-vector. Assume that $\hat{\boldsymbol{\gamma}}=(\hat{\gamma}_0^\tau,...,\hat{\gamma}_p^\tau)^\tau$ is determined by the following minimization problem:

$$\min_{\boldsymbol{\gamma}}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho\Big(y_{ij}-\gamma_0^\tau\mathbf{b}_{\mathrm{C}}(t_{ij})-\sum_{k=1}^{p}\gamma_k^\tau\mathcal{T}_k(t_{ij},x_{ijk})\Big), \tag{2.3}$$

where ρ is a given loss function and $\boldsymbol{\gamma}=(\gamma_0^\tau,...,\gamma_p^\tau)^\tau$.

For each given k, we find a point (t_{k0},x_{k0}), such that $g_k(t_{k0},x_{k0})\neq0$; then, $\xi_k(t|t_{k0})=\frac{g_k(t,x_{k0})}{g_k(t_{k0},x_{k0})}=\frac{\alpha_k(t)}{\alpha_k(t_{k0})}$ is well defined and depends on the selection of t_{k0}. Denoting $\hat{g}_k(t,x_k)=\hat{\gamma}_k^\tau\mathcal{T}_k(t,x_k)$, we approximate $\xi_k(t|t_{k0})$ as $\hat{\xi}_k(t|t_{k0},x_{k0})=\frac{\hat{g}_k(t,x_{k0})}{\hat{g}_k(t_{k0},x_{k0})}$, which depends on the selection of t_{k0} and x_{k0}. Together with the identification conditions of α_k, we obtain the spline-based initial M-estimator of $\alpha_k(k=0,...,p)$ as

$$\hat{\alpha}_{0,\mathrm{I}}(t)=\hat{\gamma}_0{}^\tau\mathbf{b}_{\mathrm{C}}(t),\qquad\hat{\alpha}_{k,\mathrm{I}}(t|t_{k0},x_{k0})=\frac{\hat{\xi}_k(t|t_{k0},x_{k0})}{\int_0^1\hat{\xi}_k(t|t_{k0},x_{k0})\mathrm{d}t}, \tag{2.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

where the subscript "I" denotes the initial estimator of α_k.

Step II: M-estimators of additive component functions

Substituting (2.4), the initial M-estimator of α_k obtained in the Step-I estimation, into VCAM (2.2), we obtain the approximated AM $y_{ij} \approx \hat{\alpha}_{0,\mathrm{I}}(t_{ij}) + \sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0})\beta_k(x_{ijk}) + \varepsilon_{ij}$, which gives a spline-based M-estimator of β_k. Denote the number of interior knots of the B-spline basis as K_{A}. Let $\boldsymbol{\theta} = (\theta_1^\tau, ..., \theta_p^\tau)^\tau$, with θ_k a $(q + K_{\mathrm{A}} - 1)$-vector, such that $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1^\tau, ..., \hat{\theta}_p^\tau)^\tau$ minimizes the following problem:

$$\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \rho\Big(y_{ij} - \hat{\alpha}_{0,\mathrm{I}}(t_{ij}) - \sum_{k=1}^{p} \hat{\alpha}_{k,\mathrm{I}}(t_{ij}|t_{k0}, x_{k0})\theta_k^\tau \mathbf{B}_{k,\mathrm{A}}(x_{ijk})\Big). \tag{2.5}$$

Then, the spline-based M-estimators $\hat{\beta}_k$, for $k = 1, ..., p$, of the additive component functions are given by

$$\hat{\beta}_k(x_k) = \check{\beta}_k(x_k) - \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i} \check{\beta}_k(x_{ijk}), \tag{2.6}$$

where $\check{\beta}_k(x_k) = \hat{\theta}_k^\tau \mathbf{B}_{k,\mathrm{A}}(x_k)$ and $N = \sum_{i=1}^{n} n_i$.

Step III: Updated M-estimators of varying-coefficient component functions

Substituting (2.6) into (2.2), we obtain an approximated VCM, $y_{ij} \approx \alpha_0(t_{ij}) + \sum_{k=1}^{p} \alpha_k(t_{ij})\hat{\beta}_k(x_{ijk}) + \varepsilon_{ij}$. Let K_{C} be the number of interior knots of the B-spline basis fitting α_k. Denote $\mathbf{h} = (h_0^\tau, ..., h_p^\tau)^\tau$, with h_k a $(q + K_{\mathrm{C}})$-vector, such that $\hat{\mathbf{h}} = (\hat{h}_0^\tau, ..., \hat{h}_p^\tau)^\tau$ minimizes

$$\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \rho\Big(y_{ij} - h_0^\tau \mathbf{b}_{\mathrm{C}}(t_{ij}) - \sum_{k=1}^{p} \hat{\beta}_k(x_{ijk}) h_k^\tau \mathbf{b}_{\mathrm{C}}(t_{ij})\Big). \tag{2.7}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

Then the updated M-estimators of α_k, for $k = 0, ..., p$, are given by

$$\hat{\alpha}_0(t) = \hat{h}_0^{\tau}\mathbf{b}_{\mathrm{C}}(t), \quad \hat{\alpha}_k(t) = \frac{\hat{h}_k^{\tau}\mathbf{b}_{\mathrm{C}}(t)}{\int_0^1 \hat{h}_k^{\tau}\mathbf{b}_{\mathrm{C}}(t)\mathrm{d}t}.$$

Common convex loss functions include the quadratic function $\rho(u) = u^2$, the check function $\rho(u) = |u| + (2\tau - 1)u$, with $\tau \in (0, 1)$, and the Huber function $\rho(u) = 0.5u^2\mathbf{I}_{|u|<\delta}$, where δ is a prespecified threshold value and $\mathbf{I}_A$ denotes the indictor function of a nonempty set A. Our method also allows for a nonconvex loss function, such as those of Hampel and Tukey. Note that the proposed estimation method has a wide range of applications, because the spline approximations in the three estimation steps are valid for both sparse data and dense data, allowing the covariates to depend simultaneously on the observation time. A simulation example given in Section S1.3 of the Supplementary Material compares our estimation method with that of Zhang and Wang (2015) when the covariates are independent of the observation time. Table 6 in the Supplementary Material shows that our estimators are superior to Zhang's estimators for sparse data and a small proportion of outliers, and perform similarly for dense data with a normal error distribution.

3. Asymptotic Results

In this section, we construct the consistency and asymptotic normality of the proposed M-estimators. Note that the asymptotic properties are considered for sparse data and dense data within a unified framework, which can be viewed as a generalization of Zhang and Wang (2016) to a VCAM. The assumptions necessary for deriving the asymptotic results are given in the Appendix.

3.1. Consistency of three-step M-estimators

Let $\bar{N}_{\mathrm{H}} = \left(\sum_{i=1}^{n} n_i^{-1}/n\right)^{-1}$ be the harmonic average of sequence $\{n_i\}$, and let $\bar{h} = \hbar_{\mathrm{A}} \vee \hbar_{\mathrm{C}}$ be the maximum of $\hbar_{\mathrm{A}}$ and $\hbar_{\mathrm{C}}$. Denote $\mathcal{J} = \{(\mathbf{x}_{ij}, t_{ij}) :, i = 1, ..., n; j = 1, ..., n_i\}$. Theorem 1 presents the rate of convergence for the additive component function β_k in the sense of the L_2-norm and the mean squared errors (MSEs).

Theorem 1. *Under Assumptions A1–A5, M1 and M2, or N1 and N2, if* $\bar{h} = O(K_{\mathrm{A}})$, $\bar{h}^2 K_{\mathrm{A}}^{2r} = o(n\bar{N}_{\mathrm{H}})$, $K_{\mathrm{A}}^2 = o(n\bar{N}_{\mathrm{A}})$, $K_{\mathrm{A}}^{2r}/n \to C_1$, $K_{\mathrm{A}}^{2r+1}/(n\bar{N}_{\mathrm{H}}) \to C_2$, *and* $K_{\mathrm{A}}/\bar{N}_{\mathrm{H}} \to C_3$, *where* $0 \le C_1 < \infty$, $0 \le C_2, C_3 \le \infty$, *then we have the convergence rates*

$$\left\|\hat{\beta}_k - \beta_k\right\|_{L_2}^2 = O_p\left(K_{\mathrm{A}}^{-2r} + \frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\right)$$

in the L_2-norm sense, and

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left[\hat{\beta}_k(x_{ijk}) - \beta_k(x_{ijk})\right]^2 = O_p\left(K_{\mathrm{A}}^{-2r} + \frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\right)$$

in the MSE sense.

Remark 2. It is easy to show the following

(i) $\frac{1}{n} = o\left(\frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}}\right)$ if $\bar{N}_{\mathrm{H}}/n^{\frac{1}{2r}} \to 0$ and $K_{\mathrm{A}} \asymp \left(n\bar{N}_{\mathrm{H}}\right)^{\frac{1}{2r+1}}$;

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

(ii) $\frac{1}{n} \asymp \frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}}$ if $\bar{N}_{\mathrm{H}}/n^{\frac{1}{2r}} \to C$ and $K_{\mathrm{A}} \asymp n^{\frac{1}{2r}}$;

(iii) $\frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}} = o(\frac{1}{n})$ if $\bar{N}_{\mathrm{H}}/n^{\frac{1}{2r}} \to \infty$ and $K_{\mathrm{A}} = o(n^{\frac{1}{2r}})$.

That is, the order of the variance term $\frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}$ has either a parametric rate of convergence $\frac{1}{n}$ or a nonparametric rate of convergence $\frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}}$, depending on the magnitude of $\bar{N}_{\mathrm{H}}/n^{\frac{1}{2r}}$.

Theorem 2 is the analogue of Theorem 1 for the varying-coefficient function α_k.

Theorem 2. *Under Assumptions A1–A5, M1 and M2, or N1 and N2, if* $K_{\mathrm{A}}K_{\mathrm{C}}^{2r} = o(n\bar{N}_{\mathrm{H}})$, $K_{\mathrm{A}} = O(K_{\mathrm{C}})$ *or* $K_{\mathrm{A}} = o(K_{\mathrm{C}})$, $K_{\mathrm{C}}^{2r}/n \to C_1$, $K_{\mathrm{C}}^{2r+1}/(n\bar{N}_{\mathrm{H}}) \to C_2$, *and* $K_{\mathrm{C}}/\bar{N}_{\mathrm{H}} \to C_3$, *where* $0 \le C_1 < \infty$, $0 \le C_2$, $C_3 \le \infty$, *then we have*

$$\|\hat{\alpha}_k - \alpha_k\|_{L_2}^2 = O_p\Big(K_{\mathrm{C}}^{-2r} + \frac{K_{\mathrm{C}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big)$$

in the L_2*-norm sense, and*

$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\big[\hat{\alpha}_k(t_{ij}) - \alpha_k(t_{ij})\big]^2 = O_p\Big(K_{\mathrm{C}}^{-2r} + \frac{K_{\mathrm{C}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big)$$

in the MSE sense.

A remark similar to that for Theorem 1 can be made for M-estimators of varying-coefficient functions. Based upon these statements, we say that the data are

- sparse if $\bar{N}_{\mathrm{H}}/n^{\frac{1}{2r}} \to 0$, which yields a nonparametric rate; or
- dense if $\bar{N}_{\mathrm{H}}/n^{\frac{1}{2r}} \to C$, with $0 < C \leq \infty$, which yields a parametric rate.

We generalize the way we split sparse data and dense data in that our conclusions reduce to those of Zhang and Wang (2016) when $r = 2$.

3.2. Asymptotic normality of three-step M-estimators

In this subsection, we present the asymptotic distribution of the M-estimators. First, we introduce the following notation:

$$W_{n,\mathrm{A}} = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \varpi(t_{ij}) \Psi_{ij} \Psi_{ij}^{\tau}, \qquad U_{n,\mathrm{A}} = \sum_{i=1}^{n} \mathbf{\Psi}_i^{\tau} G_i \mathbf{\Psi}_i / n_i^2,$$

$$W_{n,\mathrm{C}} = \sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \varpi(t_{ij}) \Phi_{ij} \Phi_{ij}^{\tau}, \qquad U_{n,\mathrm{C}} = \sum_{i=1}^{n} \mathbf{\Phi}_i^{\tau} G_i \mathbf{\Phi}_i / n_i^2,$$

where

$$\mathbf{\Psi}_i = \{\Psi_{i1}, ..., \Psi_{in_i}\}^{\tau},\ \Psi_{ij} = \left\{\psi_1(x_{ij1})^{\tau}, ..., \psi_p(x_{ijp})^{\tau}\right\}^{\tau},\ \psi_k(x_{ijk}) = \alpha_k(t_{ij}) \mathbf{B}_{k,\mathrm{A}}(x_{ijk}),$$

$$\mathbf{\Phi}_i = \{\Phi_{i1}, ..., \Phi_{in_i}\}^{\tau},\ \Phi_{ij} = \{1, \beta_1(x_{ij1}), ..., \beta_p(x_{ijp})\}^{\tau} \otimes \mathbf{b}_{\mathrm{C}}(t_{ij}).$$

Theorem 3 presents the asymptotic distribution for the additive function β_k.

Theorem 3. *Under the conditions of Theorem 1, if* $K_{\mathrm{A}}^{2r} \tilde{K}_{\mathrm{A}}/n \to \infty$,

$$\frac{\max\left(K_{\mathrm{A}}^{-3/2} \sum_{i=1}^{n} 1/n_i^2, K_{\mathrm{A}}^{1/2} \sum_{i=1}^{n} (n_i - 1)/n_i^2, \sum_{i=1}^{n} (n_i - 1)(n_i - 2)/n_i^2\right)}{\left(\sum_{i=1}^{n} \frac{1}{n_i}(K_{\mathrm{A}} - 1) + n\right)^{3/2}} \to 0,$$

and the largest eigenvalue of $K_{\mathrm{A}} \mathbf{B}_{k,\mathrm{A}}(x) \mathbf{B}_{k,\mathrm{A}}(x)^{\tau}$ *is bounded, then given* $\mathcal{J}$, *it follows that* $\hat{\beta}_k(x) - \beta_k(x) \xrightarrow{D} N(0, D_{n,\mathrm{A}}(x))$, *where*

$$D_{n,\mathrm{A}}(x) = A_k(x)^{\tau} W_{n,\mathrm{A}}^{-1} U_{n,\mathrm{A}} W_{n,\mathrm{A}}^{-1} A_k(x), \tag{3.1}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

and $A_k(x) = \{\mathbf{0}, ..., \mathbf{B}^{\tau}_{k,\mathrm{A}}(x), ...\mathbf{0}\}^{\tau}$ *is a* $\{pJ_{\mathrm{A}}\}$*-dimensional vector, with* $\mathbf{B}_{k,\mathbf{A}}(x)$ *in its* $\{(k-1)J_{\mathrm{A}}\}$*th to* $\{kJ_{\mathrm{A}}\}$*th positions, and zeros in the rest.*

Theorem 4 is the analogue of Theorem 3 for the varying-coefficient function α_k.

Theorem 4. *Under the conditions of Theorem 2, if* $K_{\mathrm{C}}^{2r}\tilde{K}_{\mathrm{C}}/n \to \infty$,

$$\frac{\max\left(K_{\mathrm{C}}^{3/2}\sum_{i=1}^{n} 1/n_i^2, K_{\mathrm{C}}^{1/2}\sum_{i=1}^{n}(n_i-1)/n_i^2, \sum_{i=1}^{n}(n_i-1)(n_i-2)/n_i^2\right)}{(\sum_{i=1}^{n}\frac{1}{n_i}(K_{\mathrm{C}}-1)+n)^{3/2}} \to 0, \quad (3.2)$$

and the largest eigenvalue of $K_{\mathrm{C}}\mathbf{b}_{\mathrm{C}}(t)\mathbf{b}_{\mathrm{C}}(t)^{\tau}$ *is bounded, then given* $\mathcal{J}$*, it follows that* $\hat{\alpha}_k(t) - \alpha_k(t) \xrightarrow{D} N(0, D_{n,\mathrm{C}}(t))$*, where*

$$D_{n,\mathrm{C}}(t) = C_k(t)^{\tau} W_{n,\mathrm{C}}^{-1} U_{n,\mathrm{C}} W_{n,\mathrm{C}}^{-1} C_k(t), \quad (3.3)$$

and $C_k(t) = \{\mathbf{0}, ..., \mathbf{b}^{\tau}_{\mathrm{C}}(t), ...\mathbf{0}\}^{\tau}$ *is a* $\{(p+1)J_{\mathrm{A}}\}$*-dimensional vector, with* $\mathbf{b}_{\mathbf{C}}(t)$ *in its* $\{kJ_{\mathrm{C}}\}$*th to* $\{(k+1)J_{\mathrm{C}}\}$*th positions, and zeros in the rest.*

Now, we build a consistent estimate for the asymptotic variance given in (3.1) and (3.3). Let $\hat{G}_i = \phi(\hat{\varepsilon}_i)\phi(\hat{\varepsilon}_i)^{\tau}$, with $\phi(\hat{\varepsilon}_i) = \{\phi(\hat{\varepsilon}_{i1}), ..., \phi(\hat{\varepsilon}_{in_i})\}^{\tau}$ and $\hat{\varepsilon}_{ij} = y_{ij} - \hat{\alpha}_0(t_{ij}) - \sum_{k=1}^{p}\hat{\alpha}_k(t_{ij})\hat{\beta}_k(x_{ijk})$. Set

$$\hat{W}_{n,\mathrm{A}} = \sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\varpi(t_{ij})\hat{\Psi}_{ij}\hat{\Psi}_{ij}^{\tau}, \qquad \hat{U}_{n,\mathrm{A}} = \sum_{i=1}^{n}\hat{\mathbf{\Psi}}_i\hat{G}_i\hat{\mathbf{\Psi}}_i/n_i^2,$$
$$\hat{W}_{n,\mathrm{C}} = \sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\varpi(t_{ij})\hat{\Phi}_{ij}\hat{\Phi}_{ij}^{\tau}, \qquad \hat{U}_{n,\mathrm{C}} = \sum_{i=1}^{n}\hat{\mathbf{\Phi}}_i\hat{G}_i\hat{\mathbf{\Phi}}_i/n_i^2,$$

where $\hat{\Psi}_{ij}$ and $\hat{\Phi}_{ij}$ are the counterparts of Ψ_{ij} and Φ_{ij}, respectively, after replacing α_k with $\hat{\alpha}_{k,\mathrm{I}}$ and replacing β_k with $\hat{\beta}_k$. Then, the natural estimates of $D_{n,\mathrm{A}}(x)$ and $D_{n,\mathrm{C}}(t)$ are

$$\hat{D}_{n,\mathrm{A}}(x) = A_k(x)^\tau \hat{W}_{n,\mathrm{A}}^{-1} \hat{U}_{n,\mathrm{A}} \hat{W}_{n,\mathrm{A}}^{-1} A_k(x) \quad \text{and}$$

$$\hat{D}_{n,\mathrm{C}}(t) = C_k(t)^\tau \hat{W}_{n,\mathrm{C}}^{-1} \hat{U}_{n,\mathrm{C}} \hat{W}_{n,\mathrm{C}}^{-1} C_k(t).$$

Theorem 5 shows that the estimates of the asymptotic variances are consistent.

Theorem 5. *Suppose that* $\sup_{t\in[0,1]} \mathrm{E}(\phi^4(\varepsilon_{ij})|t_{ij} = t) < \infty$.

(i) Under the conditions of Theorem 3, if $K_\mathrm{A} = o(\bar{h}^r)$, $K_\mathrm{A}^2 = o(n\bar{N}_\mathrm{H})$, $\max_i n_i K_\mathrm{A}^2 = o(n\bar{N}_\mathrm{H})$, *and* $K_\mathrm{A}^4 \max_i n_i = o(n^4)$, *then it holds that* $\hat{D}_{n,\mathrm{A}}(x) \xrightarrow{p} D_{n,\mathrm{A}}(x)$.

(ii) Under the conditions of Theorem 4, if $K_\mathrm{C} = o(K_\mathrm{A})$, $K_\mathrm{C}^2 = o(n\bar{N}_\mathrm{H})$, $K_\mathrm{C}^2 \max_i n_i = o(n\bar{N}_\mathrm{H})$, *and* $K_\mathrm{C}^4 \max_i n_i = o(n^4)$, *then it holds that* $\hat{D}_{n,\mathrm{C}}(x) \xrightarrow{p} D_{n,\mathrm{C}}(x)$.

In combination with Theorems 3–5, the $(1-\alpha)\%$ confidence intervals of univariate component functions are given by

$$\hat{\alpha}_k(t) \pm z_{\alpha/2}\big\{\hat{D}_{n,\mathrm{C}}(t)\big\}^{1/2} \quad \text{and} \quad \hat{\beta}_k(x) \pm z_{\alpha/2}\big\{\hat{D}_{n,\mathrm{A}}(x)\big\}^{1/2}. \qquad (3.4)$$

3.3. Quantile regression

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

Let $0 < \tau < 1$ and loss function $\rho(u) = |u| + (2\tau - 1)u$; then, the proposed M-estimators reduce to τth quantile estimates. Denote $\hat{\alpha}_{k,\tau}(t)$ and $\hat{\beta}_{k,\tau}(x)$ as the τth quantile estimates of α_k and β_k, respectively. We impose the following additional assumptions:

(Q1) $P(\varepsilon_{ij} \leq 0|\mathbf{x}_{ij}, t_{ij}) = \tau$.

(Q2) There exist positive constants c_5 and C_6 such that the conditional density function $g(x|t)$ of ε_{ij}, given $t_{ij} = t$, satisfies $|g(x|t) - g(0|t)| \leq C_6|x|$, for all $x \in [-c_5, c_5]$ and $t \in [0, 1]$, and $g(0|t)$ is bounded away from zero and infinity uniformly over $[0, 1]$.

Noting that $\rho(u)$ is convex and $\phi(u) = \rho'(u) = 2\tau I(u > 0) + 2(\tau - 1)I(u < 0)$, it is easy to show that Assumption M2 holds. If Assumption Q1 holds, then $\mathrm{E}\phi(\varepsilon_{ij}) = 0$ and Assumption M1 holds with $\varpi(t) = 2g(0|t)$. Employing Theorems 1 and 2, we obtain the following corollary.

Corollary 1. *Suppose that conditions Q1 and Q2 hold.*

- *Under the conditions of Theorem 1, we have*

$$\left\|\hat{\beta}_{k,\tau} - \beta_k\right\|_{L_2}^2 = O_p\Big(K_{\mathrm{A}}^{-2r} + \frac{K_{\mathrm{A}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big).$$

- *Under the conditions of Theorem 2, we have*

$$\|\hat{\alpha}_{k,\tau} - \alpha_k\|_{L_2}^2 = O_p\Big(K_{\mathrm{C}}^{-2r} + \frac{K_{\mathrm{C}}}{n\bar{N}_{\mathrm{H}}} + \frac{1}{n}\Big).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

Remark 3. Let $\varpi(t) = 2g(0|t)$ in $W_{n,\mathrm{A}}$ and $W_{n,\mathrm{C}}$. If conditions Q1 and Q2 hold, then we can present the asymptotic distributions of $\hat{\beta}_{k,\tau}(x)$ and $\hat{\alpha}_{k,\tau}(t)$ under the conditions of Theorems 3 and 4, respectively.

4. Model Identification Procedure

The VCAM (2.1) is a flexible nonparametric regression method. However, parsimony is always preferable when several potential options are available. To this end, we propose a model identification strategy based on the penalized M-estimators for identifying additive terms and varying-coefficient terms.

The assumption of continuous covariates means that $X_k \neq 0$ almost surely, for $k = 1, ..., p$, and model (2.2) can be rewritten as

$$y_{ij} = \alpha_0(t_{ij}) + \sum_{k=1}^{p} x_{ijk}\alpha_k(t_{ij})\beta_k^*(x_{ijk}) + \varepsilon_{ij},$$

where $\beta_k^*(x) = \beta_k(x)/x$. Employing the tensor product of B-spline bases, the bivariate function $g_k^*(t, x_k) = \alpha_k(t)\beta_k^*(x_k)$ can be approximated as

$$\begin{aligned} g_k^*(t, x_k) \approx\ & \{1, \mathbf{B}_{k,\mathrm{AP}}^{\tau}(x_k)\} \otimes \{1, \mathbf{B}_{\mathrm{CP}}^{\tau}(t)\}\eta_k \\ &= \eta_{00,k} + \eta_{\cdot 0,k}^{\tau}\mathbf{B}_{\mathrm{CP}}(t) + \eta_{01,k}B_{k1}(x_k) + \eta_{\cdot 1,k}^{\tau}B_{k1}(x_k) \otimes \mathbf{B}_{\mathrm{CP}}(t) \\ &\quad + \cdots + \eta_{0J_{\mathrm{AP}},k}B_{kJ_{\mathrm{AP}}}(x_k) + \eta_{\cdot J_{\mathrm{AP}},k}^{\tau}B_{k,J_{\mathrm{AP}}}(x_k) \otimes \mathbf{B}_{\mathrm{CP}}(t), \end{aligned}$$

where $\eta_k = \left\{\eta_{00,k}, \eta_{\cdot 0,k}^{\tau}, \eta_{01,k}, \eta_{\cdot 1,k}^{\tau}, ..., \eta_{0J_{\mathrm{AP}},k}, \eta_{\cdot J_{\mathrm{AP}},k}^{\tau}\right\}^{\tau}$, $\eta_{\cdot j,k} = \left\{\eta_{1j,k}, ..., \eta_{J_{\mathrm{CP}}j,k}\right\}^{\tau}$, and J_{AP} and J_{CP} are the cardinalities of the B-spline bases $\mathbf{B}_{k,\mathrm{AP}}(x_k)$ and $\mathbf{B}_{\mathrm{CP}}(t)$, respectively, for β_k and α_k, in the model identification procedure.

Let $M_k(t, x_k) = \{0, \mathbf{B}_{\mathrm{CP}}^{\tau}(t), 0, B_{k1}(x_k) \otimes \mathbf{B}_{\mathrm{CP}}^{\tau}(t), ..., 0, B_{kJ_{\mathrm{AP}}}(x_k) \otimes \mathbf{B}_{\mathrm{CP}}^{\tau}(t)\}^{\tau}$ and $F_k(t, x_k) = \{0_{J_{\mathrm{CP}}+1}^{\tau}, \mathbf{B}_{k,\mathrm{AP}}^{\tau}(x_k) \otimes (1, \mathbf{B}_{\mathrm{CP}}^{\tau}(t))\}^{\tau}$, where 0_l denotes the l-vector of zeros. Then, we can say that g_k reduces to

- an additive term if and only if $\eta_k^{\tau} M_k(t, x_k) = 0$, and
- a varying-coefficient term if and only if $\eta_k^{\tau} F_k(t, x_k) = 0$,

for any $(t, x) \in [0, 1] \times [a_k, b_k]$, where $[a_k, b_k]$ is the domain of $\beta_k(\cdot)$.

We now propose a regularized M-estimation method in which we penalize the L_2-norm of $M_k^{\tau}\eta_k$ and $F_k^{\tau}\eta_k$, for $k = 1, ..., p$. Denote the numbers of interior knots for α_k and β_k in the model identification procedure as $\hbar_{\mathrm{CP}}$ and $\hbar_{\mathrm{AP}}$, respectively. Let $\boldsymbol{\eta} = (\eta_0^{\tau}, ..., \eta_p^{\tau})^{\tau}$, where η_0 is a $\{q + \hbar_{\mathrm{CP}}\}$-vector and $\eta_k (k = 1, ..., p)$ is a $\{(q + \hbar_{\mathrm{CP}})(q + \hbar_{\mathrm{AP}} - 1)\}$-vector. Suppose $\hat{\boldsymbol{\eta}} = (\hat{\eta}_0^{\tau}, ..., \hat{\eta}_p^{\tau})^{\tau}$ minimizes the following problem:

$$\begin{aligned} &\sum_{i=1}^{n} \frac{1}{n_i} \sum_{j=1}^{n_i} \rho\Big(y_{ij} - \eta_0^{\tau}\mathbf{b}_{\mathrm{C}}(t_{ij}) - \sum_{k=1}^{p} x_{ijk}\{1, \mathbf{B}_{k,\mathrm{AP}}^{\tau}(x_k)\} \otimes \{1, \mathbf{B}_{\mathrm{CP}}^{\tau}(t)\}\eta_k\Big) \\ &\qquad + n\sum_{k=1}^{p} p_{\lambda_1}(\|M_k^{\tau}\eta_k\|_{L_2}) + n\sum_{k=1}^{p} p_{\lambda_2}(\|F_k^{\tau}\eta_k\|_{L_2}). \end{aligned} \tag{4.1}$$

The product term $\alpha_k(t)\beta_k(x_k)$ in (1.1) then becomes an additive term if $\|M_k^{\tau}\hat{\eta}_k\|_{L_2}$ is close to zero (e.g., no larger than 10^{-4}), and becomes a varying-coefficient term if $\|F_k^{\tau}\hat{\eta}_k\|_{L_2}$ is close to zero.

There are various ways to specify the penalty function $p_{\lambda}(\cdot)$ (Tibshirani , 1996; Fan and Li , 2001; Zou , 2006). We adopt the smoothly clipped absolute

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

deviation (SCAD) penalty function, and use the locally quadratic approximation (LQA) algorithm proposed by Fan and Li (2001).

Let $\mathcal{I}_{\mathrm{A}}$ and $\mathcal{I}_{\mathrm{V}}$ be the index sets of additive terms and varying-coefficient terms, respectively, in VCAM (2.1). Denote $\varrho_n = \hbar_{\mathrm{P}}^{-r} + \sqrt{\kappa_{\mathrm{P}}/n}$, with $\hbar_{\mathrm{P}} = \hbar_{\mathrm{AP}} \wedge \hbar_{\mathrm{CP}}$ and $\kappa_{\mathrm{P}} = \hbar_{\mathrm{P}}^2/\bar{N}_{\mathrm{H}}$.

Theorem 6 demonstrates the consistency of the model identification procedure.

Theorem 6. *Suppose that Assumptions A1–A5, M1 and M2, or N1 and N2 hold.*

(i) If $\lambda_1 \to 0$, $\sqrt{\varrho_n}/\lambda_1 \to 0$, *and* $\liminf_{n\to\infty} \liminf_{w\to 0+} p'_{\lambda_1}(w)/\lambda_1 = 1$, *then* $M_k^{\tau}(t, x_k)\hat{\eta}_k = 0\ \forall k \in \mathcal{I}_{\mathrm{A}}$ *with probability approaching unity.*

(ii) If $\lambda_2 \to 0$, $\sqrt{\varrho_n}/\lambda_2 \to 0$, *and* $\liminf_{n\to\infty} \liminf_{w\to 0+} p'_{\lambda_2}(w)/\lambda_2 = 1$, *then* $F_k^{\tau}(t, x_k)\hat{\eta}_k = 0\ \forall k \in \mathcal{I}_{\mathrm{V}}$ *with probability approaching unity.*

5. Implementation Issues

In this section, we address several practical problems related to the selection of smoothing parameters and tuning parameters in our methods. As is common practice in the spline literature, we select the number of interior knots using a data-driven method (i.e., the Bayes information criterion; BIC), and position the knots at equal intervals on the sample quantiles.

- Selecting the optimal number of interior knots $(\hbar_{\mathrm{C}}, \hbar_{\mathrm{A}})$.

  The optimal number of interior knots $(\hat{\hbar}_{\mathrm{C}}, \hat{\hbar}_{\mathrm{A}})$ in the Step-I estimation minimizes the following BIC function:

$$\mathrm{BIC}_1(\hbar_{\mathrm{C}}, \hbar_{\mathrm{A}}) = \log\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho(\hat{\sigma}_{ij,1})\Big) + \frac{\log N}{2N}N_1,$$

  where $\hat{\sigma}_{ij,1} = y_{ij} - \hat{\gamma}_0^{\tau}\mathbf{b}_{\mathrm{C}}(t_{ij}) - \sum_{k=1}^{p}\hat{\gamma}_k^{\tau}\mathcal{T}_k(t_{ij}, x_{ijk})$ and $N_1 = (q+\hbar_{\mathrm{C}})(1+p(q+\hbar_{\mathrm{A}}-1))$.

- Selecting the optimal number of interior knots $(K_{\mathrm{A}}, K_{\mathrm{C}})$.

  The optimal number of interior knots $(\hat{K}_{\mathrm{A}}, \hat{K}_{\mathrm{C}})$ in Steps II and III minimizes

$$\mathrm{BIC}_2(K_{\mathrm{A}}, K_{\mathrm{C}}) = \log\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho(\hat{\sigma}_{ij,2})\Big) + \frac{\log N}{2N}N_2,$$

  where $\hat{\sigma}_{ij,2} = y_{ij} - \hat{\alpha}_0(t_{ij}) - \sum_{k=1}^{p}\hat{\alpha}_k(t_{ij})\hat{\beta}_k(x_{ijk})$ and $N_2 = p(q+K_{\mathrm{A}}-1) + (p+1)(q+K_{\mathrm{C}})$.

- Selecting the optimal tuning parameters (λ_1, λ_2).

  We use the optimal number of interior knots $(\hat{\hbar}_{\mathrm{C}}, \hat{\hbar}_{\mathrm{A}})$ and the optimal tuning parameters $(\hat{\lambda}_1, \hat{\lambda}_2)$ that minimize the following BIC:

$$\mathrm{BIC}_3(\lambda_1, \lambda_2) = \log\Big(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\rho(\hat{\sigma}_{ij,3})\Big) + \frac{\log N}{2N}N_3,$$

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

where $\hat{\sigma}_{ij,3} = y_{ij} - \hat{\eta}_0^\tau \mathbf{b}_{\mathrm{CP}}(t_{ij}) - \sum_{k=1}^{p} x_{ijk}\{1, \mathbf{B}_{k,\mathrm{AP}}^\tau(x_{ijk})\} \otimes \{1, \mathbf{B}_{\mathrm{CP}}^\tau(t_{ij})\}\hat{\eta}_k$ and $N_3 = m_{\mathrm{L}} + \{q + \hat{\hbar}_C\}\{m_{\mathrm{C}} + 1\} + m_{\mathrm{A}}\{q + \hat{\hbar}_A - 1\} + \{q + \hat{\hbar}_C\}\{q + \hat{\hbar}_A - 1\}\{p - m_{\mathrm{L}} - m_{\mathrm{C}} - m_{\mathrm{A}}\}$, with m_{L} linear terms, m_{A} additive terms, and m_{C} varying-coefficient terms.

6. Numerical Studies

Simulation examples are used to investigate the finite-sample performance of the proposed three-step M-estimation method and model identification procedure. Empirical examples are then presented to illustrate the usefulness of our method in practice.

6.1. Simulation studies

Example 1. A VCAM with repeated measurements is generated as follows:

$$y_{ij} = \alpha_0(t_{ij}) + \alpha_1(t_{ij})\beta_1(x_{ij}) + w_i(t_{ij}) + e_{ij}, \quad i = 1, ..., n; \ j = 1..., m,$$

where t_{ij} are independent and identically distributed (i.i.d.) copies from $U(0,1)$, and $x_{ij} = 0.8t_{ij}^2 + \eta_{ij}$, with η_{ij} drawn independently from $N\big(0, (1+t_{ij})/(2+t_{ij})\big)$. The subject-specific random trajectories $w_i(i = 1, ..., n)$ are independent copies of a zero-mean stationary Gaussian process with covariance function $\gamma(u) = 0.35\theta^{|u|}$, where $\theta = 0$ and 0.5. The random noise e_{ij} are i.i.d. from four error distributions: the normal distribution $N(0, 0.2)$, the mixed normal distribution $0.95N(0, 0.2) + 0.05N(0, 12.5^2)$, and the scaled t distributions of $0.5 \times t(2)$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

$0.2 \times t(1)$. The univariate component functions are given by $\alpha_0(t) = \cos(2\pi t)$, $\alpha_1(t) = \{2t\sin(2\pi t) + 1\}/\int_0^1\{2t\sin(2\pi t) + 1\}\mathrm{d}t$, and $\beta_1(x) = 1.5\sin(\pi x/2) - x(1-x) - \mathrm{E}[1.5\sin(\pi X/2) - X(1-X)]$.

Three loss functions are considered: the quadratic function $\rho_1(x) = x^2$, the absolute value function $\rho_2(x) = |x|$, and the Huber function $\rho_3(x) = 0.5x^2\mathbf{I}_{|x|<\delta}$, with $\delta = 1.345$. We evaluate the performance of the three-step M-estimator using the MSE, which is defined as

$$\mathrm{MSE}(g) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[\hat{g}(t_{ij}) - g(t_{ij})\right]^2,$$

where g is either α_k or β_k. To obtain an intuitive impression of the robustness of the M-estimators, we define the weighted average squared error (WASE) as

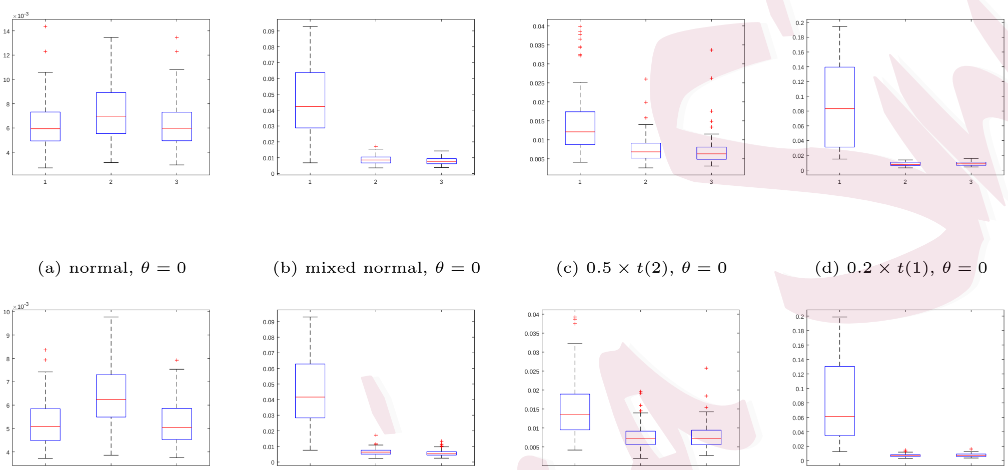
$$\mathrm{WASE} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{[\hat{\alpha}_0(t_{ij}) - \alpha_0(t_{ij})]^2}{[\mathrm{range}(\alpha_0)]^2} + \frac{[\hat{\alpha}_1(t_{ij}) - \alpha_1(t_{ij})]^2}{[\mathrm{range}(\alpha_1)]^2} + \frac{[\hat{\beta}_1(x_{ij}) - \beta_1(x_{ij})]^2}{[\mathrm{range}(\beta_1)]^2}\right\},$$

where $\mathrm{range}(f)$ denotes the range of a given function f.

For $n = 30$ and $m = 20$, based upon 500 Monte Carlo replications, Figure 1 shows the average WASE of the three-step M-estimators with the four error distributions and two types of intra-subject covariance structure. In this figure, 1, 2, and 3 denote the least-squares estimator, median estimator, and Huber estimator, respectively. We also compare the average MSE (AMSE) in Table 1 of the Supplementary Material. The results show that the Huber estimator and the median estimator perform similarly, regardless of which error distribution is adopted. In terms of performance, they are comparable to the least-squares

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

estimators under normal error distributions, and are superior to the least-squares estimators under nonnormal error distributions. In addition, the influence of the intra-subject covariance structure is nonsignificant.



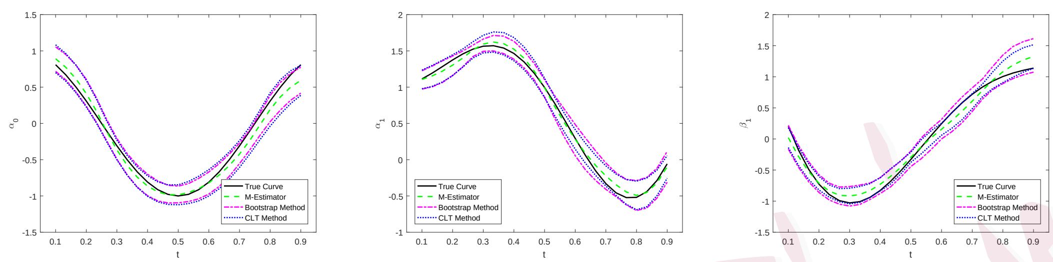
(a) normal, $\theta = 0$ (b) mixed normal, $\theta = 0$ (c) $0.5 \times t(2)$, $\theta = 0$ (d) $0.2 \times t(1)$, $\theta = 0$

(e) normal, $\theta = 0.5$ (f) mixed normal, $\theta = 0.5$ (g) $0.5 \times t(2)$, $\theta = 0.5$ (h) $0.2 \times t(1)$, $\theta = 0.5$

Figure 1: Box plot for the average WASE (AWASE) based on 500 Monte Carlo replications; 1, 2, and 3 denote least-squares estimator, median estimator, and Huber estimator, respectively.

Furthermore, we provide a graphical representation of the iterative Huber estimator under a mixed normal error distribution. Figure 2 shows the pointwise 95% confidence intervals (CIs) of the Huber estimator based on the central limit theorem (CLT) (dotted lines), and the 95% CIs based on 500 wild bootstrap samplings (dash-dotted lines). The true component function (solid line) and Huber

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483



(a) α_0 (b) α_1 (c) β_1

Figure 2: Three-step M-estimators under mixed normal error distribution. Solid line: true component function; dashed line: three-step M-estimator; dotted lines: 95% CIs based on (3.4); dash-dotted lines: 95% CIs based on 500 wild bootstrap resamplings.

M-estimator (dashed line) are also given. The figures show that the two types of CIs are not significantly different, which motivates our claim that the bootstrap method is sound. However, we do not investigate the theoretical justification for that claim to avoid straying from the primary aim of this study. However, note that the true curves and the Huber estimators are very close, and both fall into the 95% CIs, indicating the rationality of the proposed estimation method. Under a normal error distribution, the least-squares-based CIs are shown in Figure 1 of the Supplementary Material.

We also investigate the average experience coverage probability (AECP) of the three-step M-estimator with a normal error distribution and a mixed normal

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

error distribution in Figures 2 and 3, respectively, of the Supplementary Material, which show that the pointwise CLT-based CI performs well, even in the presence of a small proportion of outliers. In addition, Figure 4 in the Supplementary Material compares the AECP of the component functions under a more general sampling plan, namely, that of sparse observations for some subjects, and dense observations for other subjects. The results show that the more general sampling plan and a small proportion of outliers have no significant influence on the AECP of the component functions.

Tables 2 and 3 in the Supplementary Material also compare the average of MSE (AMSEs) of the iterative M-estimator under different combinations of (n, m), with $n = 20, 40$ and $m = 20, 30$. We conclude that, as the total number of observations grows, the AMSE decreases for a normal error distribution, regardless of which loss function is used. For nonnormal error distributions, the AMSEs of the estimators based on the loss functions ρ_2 and ρ_3 decrease, but the least-squares estimator shows no significant improvement as the total observation size grows.

The numerical example considered in Section S1.2 of the Supplementary Material investigates the finite-sample performance of the model identification procedure. As expected, the results given in Tables 4 and 5 of the Supplementary Material verify our asymptotic theories and demonstrate the robustness of the model identification.

6.2. Analysis of real data

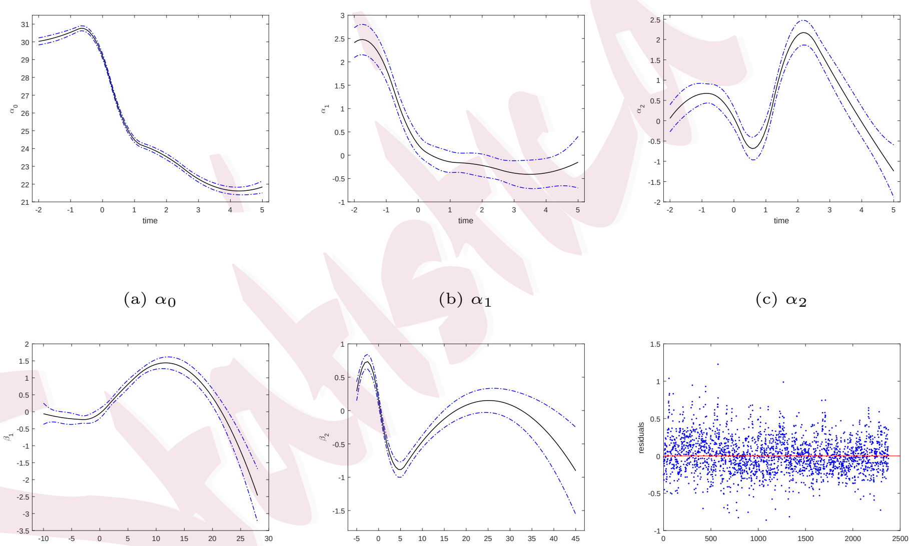
Example 2. We now apply our method to CD4 data from the Multicenter AIDS Cohort Study, which contain 2,376 observations from 369 men infected with HIV. Zhang, Park, and Wang (2013) analyzed this data set using the time-varying AM $y_{ij} = \mu_0(t_{ij}) + \sum_{k=1}^{2} \mu_k(t_{ij}, x_{ijk}) + w_{ij} + e_{ij}$. Following their work, we choose two covariates: X_1 (age), the age at seroconversion (time-invariant variable); and X_2 (cesd), the level of depression, which is recorded over time (in years).

Employing the separability test proposed by Hu, Huang, and You (2018), we obtain a p-value of 0.84, which means the VCAM (2.2), a submodel of the time-varying AM introduced by Zhang, Park, and Wang (2013), is sufficient for this data set. Under loss function ρ_3 in Example 1, we select optimal knots $(\hat{\hbar}_{\mathrm{C}}, \hat{\hbar}_{\mathrm{A}}, \hat{K}_{\mathrm{C}}, \hat{K}_{\mathrm{A}}) = (2, 2, 4, 3)$ using the BIC given in Section 5. Then, we obtain the optimal tuning parameters $(\hat{\lambda}_1, \hat{\lambda}_2) = (3.06, 1.56)$, which are selected from $[0.01, 5]$, with spacing 0.05. Based on the resulting optimal parameters, we obtain the penalized estimators. Thus, we conclude that α_1 and α_2 are time-variant and that β_1 and β_2 are nonlinear.

The Huber estimator and the 95% CIs of the univariate component functions are presented in Figure 3, from which we conclude that the overall mean functions α_0 and α_1 for X_1 (age) are monotonically decreasing, and that α_2 for X_2 (cesd)

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

is a bimodal function. For a fixed time, the effect of age on the CD4 count increases until around age = 12, after which it decreases. However, the effect of depression on the CD4 count decreases rapidly before cesd = 5, then increases until around cesd = 25, after which it decreases. The plot of the residuals in Figure 3(f) shows that our regression method is appropriate for this data set. Figure 6 in the Supplementary Material shows the estimated surfaces of the bivariate function $g_k(t, x_k) = \alpha_k(t)\beta_k(x_k)$, for $k = 1, 2$.



(a) α_0 (b) α_1 (c) α_2

(d) β_1 (e) β_2 (f) Residual plot

Figure 3: Three-step M-estimators for CD4 data set. Solid line: three-step M-estimators; dash-dotted lines: 95% CIs based on (3.4); (f) plots the scaled residuals relative to the fitted values.

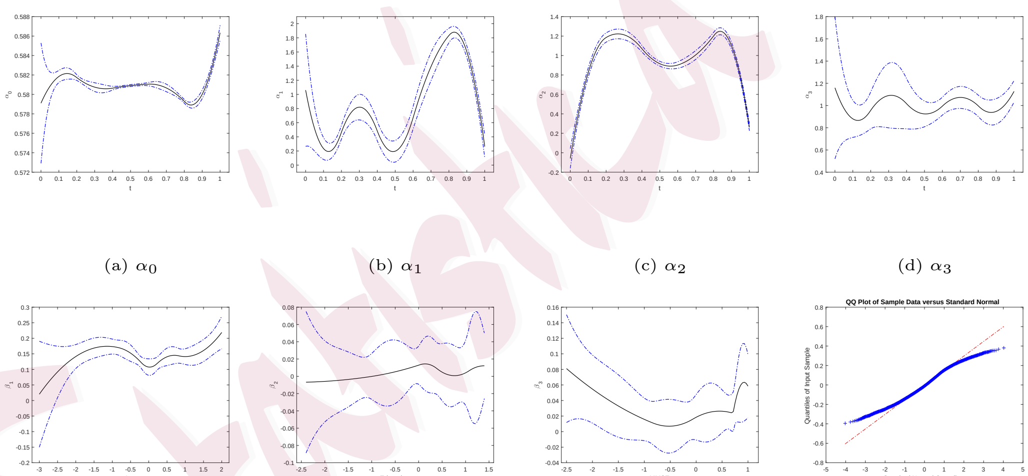
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

Example 3. In this example, we consider a real diffusion-weighted imaging data set, with $n = 213$ subjects collected from the NIH Alzheimer's Disease Neuroimaging Initiative (ADNI) study. The observed response process is a fractional anisotropy (FA) curve at all 83 grid points along the skeleton of the midsagittal corpus callosum. Here, we want to explore the relationship between FA (Y) and three covariates: (i) the age of the subject (X_1); (ii) their educational level (X_2); and (iii) the result of the ADNI Mini-Mental State Exam (X_3). Luo, Zhu, and Zhu (2016) and Li, Huang, and Zhu (2017) analyzed this data set using a single-index VCM and a functional varying-coefficient single-index model, respectively. The two models both assume linear covariate effects with varying coefficients and/or nonlinear covariate effects only through the linear combination of the covariates with varying coefficients. However, the linear effect is a somewhat strict constraint in practical applications. Furthermore, we are interested in the function effect of each predictor on the response process, including the linear effect as its special case. Therefore, we apply a VCAM to this data set.

Employing the proposed model identification procedure, we claim that the varying-coefficient functions are all time-variant, and that the additive functions are all nonlinear. Figure 4 shows the Huber estimators of the univariate component functions and the 95% pointwise CIs based on (3.4). Figure 4(e)–(g) show how the covariates affect the response process: the effect of age increases

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

initially, then decreases before the average age, and subsequently increases; the effect of educational level increases gently before the average educational level, then decreases, and finally increases; the effect of the ADNI Mini-Mental State Exam decreases until nearly the average value, and then increases. The estimated bivariate functions $g_k(t, x_k) = \alpha_k(t)\beta_k(x_k)$, for $k = 1, 2, 3$, are presented in Figure 10 of the Supplementary Material, which shows the dynamic effects of the covariates. The Q–Q plot shows that our regression method is appropriate for this data set.



(a) α_0 (b) α_1 (c) α_2 (d) α_3

(e) β_1 (f) β_2 (g) β_3 (h) Q-Q plot

Figure 4: Estimated univariate component functions for ADNI data. Solid line: three-step M-estimator; dash-dotted lines: 95% CIs based on (3.4).

An analysis of the cigarette data mentioned in Section 1 shows that a reduced

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

VCAM is preferable. Details are given in Section S1.6 of the Supplementary Material.

7. Conclusion

The VCAM proposed by Zhang and Wang (2015) is a flexible structural nonparametric regression method that includes the classical VCM and AM as special cases. In this study, we developed an M-type robust regression method for this VCAM to enable analyses of longitudinal data and functional data, which may include sparse or dense repeated measurements for the selected subjects, and both the response and the covariates may be smooth processes that depend on the observation time.

We have proposed spline-based three-step M-estimators for varying-coefficient component functions and additive component functions. The asymptotic properties are considered for sparse and dense data within a unified framework, which separates these data based on the relative order of n_i to n. Similarly to Hu, Huang, and You (2018), the proposed estimation method exhibits the oracle property in that the iterative estimation procedure does not cause additional asymptotic errors.

To select as parsimonious a model as possible, we have also developed a model identification procedure based on the SCAD penalty function. Here, we showed that the proposed model identification method correctly selects an additive term and a varying-coefficient term with probability approaching unity.

Acknowledgments

The authors are grateful to the co-editor, an associate editor, and the referees for their constructive comments. Dr. Huang's research was partially supported by National Natural Science Foundation of China (NSFC) (No.11871323) and the State Key Program in the Major Research Plan of NSFC (No. 91546202). Dr. You's research was partially supported by National Natural Science Foundation of China (NSFC) (No.11971291) and the Program for Innovative Research Team of SHUFE.

Appendix

Let $C_r[a,b]$ be the space of all functions $m(x)$ defined on $[a,b]$ such that the $(r-1)$-order derivative $m^{(r-1)}(\cdot)$ is continuous over $[a,b]$, and

$$\left|m^{(r-1)}(x) - m^{(r-1)}(x')\right| \leq C|x-x'|, \quad \forall\, x, x' \in [a,b],$$

where C is a positive constant. The necessary conditions for the asymptotic results are listed below.

- Basic assumptions.

(A1) The time points $\{t_{ij}\}$ are independent copies of T, whose probability density function $f_T(\cdot)$ is uniformly bounded away from zero and infinity.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

(A2) The marginal density function $f_k(\cdot)$ of X_k is uniformly bounded away from zero and infinity over the support set S_k of X_k. The joint density $f_{\mathbf{X},T}(\mathbf{x},t)$ of $\mathbf{X}$ and T is uniformly bounded away from zero and infinity on $(\mathbf{x},t) \in \prod_{k=1}^{p} S_k \times [0,1]$.

(A3) $\alpha_k \in C_r[0,1]$ and $\beta_k \in C_r[a_k, b_k]$, where $1 \le r \le q$ and a_k, b_k are finite real numbers for $k = 1, ..., p$.

(A4) The function $\phi(\cdot) = \rho'(\cdot)$ satisfies $\mathrm{E}[\phi(\varepsilon_{ij})|t_{ij} = t] = 0$ and $\mathrm{E}[\phi^2(\varepsilon_{ij})|t_{ij} = t] \le C_1$ for any $t \in [0,1]$, where C_1 is a positive constant.

(A5) There exists some positive constant $\tilde{\lambda}$ such that the smallest eigenvalue λ_{i1} of $G_i = \mathrm{E}[\phi(\varepsilon_i)\phi(\varepsilon_i)^{\tau}|\mathcal{J}]$ satisfies $\lambda_{i1} \ge \tilde{\lambda} > 0$.

- Assumptions for convex loss function.

(M1) The loss function $\rho(\cdot)$ is convex, and there exist some function $\varpi(t)$ and positive constants c_1 and C_2 such that

$$|\mathrm{E}[\phi(\varepsilon_{ij} + u)|t_{ij} = t] - \varpi(t)u| \le C_2 u^2$$

for any $|u| \le c_1$ and $t \in [0,1]$. Moreover, $\varpi(t)$ satisfies $0 < c_{\varpi} \le \min_{t\in[0,1]} \varpi(t) \le \max_{t\in[0,1]} \varpi(t) \le C_{\varpi} < \infty$.

(M2) There exist positive finite constants c_2, C_3, and C_4 such that

$$\mathrm{E}[\{\phi(\varepsilon_{ij} + u) - \phi(\varepsilon_{ij})\}^2|\mathcal{J}] \le C_3|u|$$

and $|\phi(u+v)-\phi(v)| \leq C_4$ for any $|u| \leq c_2$, $t \in [0,1]$, and $v \in \mathbb{R}$.

- Assumptions for non-convex loss function.

(N1) The function $\phi(\cdot)$ is continuous and has a derivative $\phi'(\cdot)$ almost everywhere. Furthermore, $\phi_\varepsilon(t) = \mathrm{E}[\phi'(\varepsilon_{ij})|t_{ij} = t]$ is positive and continuous at t.

(N2) $\mathrm{E}\big[\sup_{\|z\|\leq\delta} |\phi(\varepsilon_{ij}+z) - \phi(\varepsilon_{ij}) - \phi'(\varepsilon_{ij})z|t_{ij} = t\big] = o(\delta)$ as $\delta \to 0$.

Remark 4. Assumptions A1 and A2 relate to the distributions of time points t_{ij} and covariates $\mathbf{x}_{ij}$. Assumption A3 specifies the degree of smoothness of varying-coefficient component functions and additive component functions. Assumptions A4, M1, and M2 are standard assumptions about the score function ϕ of a convex loss function; see He, Zhu, and Fung (2002); Tang and Cheng (2008) for details. Assumptions N1 and N2 are necessary for a non-convex loss function; see Fan and Jiang (2000); Jiang and Mack (2001).

Supplementary Material

The online Supplementary Material includes additional numerical studies, an iterative algorithm for penalized M-estimators, and proofs of the asymptotic results.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

References

Bada, O. and Liebl D. (2012). phtt: Panel data analysis with heterogeneous time trends. R package version 3 (2).

Carroll, R. J., Maity, A., Mammen, E., and Yu, K. (2009). Nonparametric additive regression for repeatedly measured data. *Biometrika*, **96**, 383–398.

De Boor. (1978). *A Practical Guide to Splines*. Springer, New York.

Diggle P. J., Liang, K. Y., and Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford, U.K.: Oxford University Press.

Fan, J. and Jiang, J. (2000). Variable bandwidth and one-step local M-estimator. *Science in China*, **43**,65–81.

Fan, J., and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348–1360.

He, X., and Shi, P. (1994). Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journal of Nonparametric Statistics*, **3**, 299–308.

He, X., Zhu, Z., and Fung, W. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika*, **89**, 579–590.

Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Hu, L. X., Huang, T., and You J. H. (2018). Estimation and Identification of a Varying-Coefficient AdditiveModel for Locally Stationary Processes. *Journal of the American Statistical Association*, doi: 10.1080/01621459.2018.1482753.

Jiang, J. and Mack Y. (2001). Robust local polynomial regression for dependent data. *Statistica Sinica*, **11**, 705–722.

Koenker, R., and Bassett, G. (1978). Regression quantiles. *Econometrica.* **46**, 33–50.

Li, Y. H., and Hsing, T. (2010). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, **38**, 3321 – 3351.

Li, J., Huang, C., Zhu, H., & for the Alzheimers Disease Neuroimaging Initiative (2017). A Functional Varying-Coefficient Single-Index Model for Functional Response Data. *Journal of the American Statistical Association*, **112**, 1169-1181, doi: 10.1080/01621459.2016.1195742.

Luo, X., Zhu, L., and Zhu, H. (2016). Single-index varying coefficient model for functional responses. *Biometrics*, **72**, 1275–1284.

Ramsay, J. O., and Ramsey, J. B. (2002) Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Economics*, **107**, 327–344.

Tang, Q., and Cheng, L. (2008). M-estimation and B-spline approximation for varying coefficient models with longitudinal data. *Journal of Nonparametric Statistics*, **20**, 611–625.

Tibshirani, R. (1996). Regression shrinkage and selection vis the Lasso. *Journal of the Royal*

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0483

Statistical Society. Series B (Methodological), **58**, 267–288.

Vogt, M. (2012). Nonparametric regression for locally stationary time series. The Annals of Statistics, 40(5), 2601–2633.

Wang, L., and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics*, **35**, 2474–2503.

Xue, L., Qu, A., and Zhou, J. (2010). Consistent model selection for marginal generalized additive model for correlated data. *Journal of the American Statistical Association*, **105**, 1518–1530.

Xue, L., and Zhu, L. (2007). Empirical likelihood for a varying coefficient model with longitudinal data. *Journal of the American Statistical Association*, **102**, 642–654.

Yao, F. (2007). Asymptotic distributions of nonparametric regression estimators for longitudinal or functional data. *Journal of Multivariate Analysis*, **98**, 40 –56.

Yao, F., Müller, H. G., and Wang, J. L. (2005). Functional linear regression analysis for longitudinal data. *Annals of Statistics*, **33**, 2873–2903.

Zeger, S. L. and Diggle, P. G. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.

Zhang, X., Park, B. U., and Wang J. (2013). Time-varying additive models for longitudinal data. *Journal of the American Statistical Association*, **108**, 983–998.

Zhang, X., and Wang, J.-L. (2015). Varying-coefficient additive models for functional data.

Biometrika, **102**, 15–32.

Zhang, X., and Wang, J.-L. (2016). From sparse to dense functional data and beyond. *The Annals of Statistics*, **44**, 2281–2321.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association,* **101**, 1418–1429.

Lixia Hu

School of Statistics and Mathematics, Shanghai Lixin University of Accounting and Finance, Shanghai, China

E-mail: hulx18sufe@163.com

Tao Huang

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

E-mail: huang.tao@mail.shufe.edu.cn

Jinhong You

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

E-mail: johnyou07@163.com