Open Journal of Statistics, 2014, 4, 868-878
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ojs
http://dx.doi.org/10.4236/ojs.2014.410082

Model Detection for Additive Models with Longitudinal Data

Jian Wu, Liugen Xue

[1]College of Applied Sciences, Beijing University of Technology, Beijing, China
[2]College of Science, Northeastern University, Shenyang, China
Email: wujian@emails.bjut.edu.cn, mbaron@utdallas.edu

Received 1 October 2014; revised 28 October 2014; accepted 15 November 2014

Abstract

In this paper, we consider the problem of variable selection and model detection in additive models with longitudinal data. Our approach is based on spline approximation for the components aided by two Smoothly Clipped Absolute Deviation (SCAD) penalty terms. It can perform model selection (finding both zero and linear components) and estimation simultaneously. With appropriate selection of the tuning parameters, we show that the proposed procedure is consistent in both variable selection and linear components selection. Besides, being theoretically justified, the proposed method is easy to understand and straightforward to implement. Extensive simulation studies as well as a real dataset are used to illustrate the performances.

Keywords

Additive Model, Model Detection, Variable Selection, SCAD Penalty

1. Introduction

Longitudinal data arise frequently in biological and economic applications. The challenge in analyzing longitudinal data is that the likelihood function is difficult to specify or formulate for non-normal responses with large cluster size. To allow richer and more flexible model structures, an effective semi-parametric regression tool is the additive model introduced by [1], which stipulates that

$$Y = \mu + \sum_{l=1}^{d} m_l\left(X^{(l)}\right) + \varepsilon, \tag{1}$$

where Y is a varaible of interest and $X = \left(X^{(1)}, \cdots, X^{(d)}\right)^T$ is a vector of predictor variables, μ is a

How to cite this paper: Wu, J. and Xue, L.G. (2014) Model Detection for Additive Models with Longitudinal Data. *Open Journal of Statistics*, **4**, 868-878. http://dx.doi.org/10.4236/ojs.2014.410082

unknown constant, and $m(X)=\sum_{l=1}^{d} m_l\left(X^{(l)}\right)$ are unknown nonparametric functions. As in most work on nonparametric smoothing, estimation of the non-parametric functions $m(X)=\sum_{l=1}^{d} m_l\left(X^{(l)}\right)$ is conducted on a compact support. Without loss of generality, let the compact set be $\mathcal{X}=[0,1]^d$ and also impose the condition $E\left[m_l\left(X^{(l)}\right)\right]=0$ which is required for identifiability of model (1.1), $l=1,\cdots,d$. We propose a penalized method for variable selection and model detection in model (1.1) and show that the proposed method can correctly select the nonzero components with probability approaching one as the sample size goes to infinity.

Statistical inference of additive models with longitudinal data has also been considered by some authors. By extending the generalized estimating equations approach, [2] studied the estimation of additive model with longitudinal data. [3] focuses on a nonparametric additive time-varying regression model for longitudinal data. [4] considered the generalized additive model when responses from the same cluster are correlated. However, in semiparametric regression modeling, it is generally difficult to determine which covariates should enter as nonparametric components and which should enter as linear components. The commonly adopted strategy in practice is just to consider continuous entering as nonparametric components and discrete covariates entering as parametric. Traditional method uses hypothesis testing to identify the linear and zero component. But this might be cumbersome to perform in practice whether there are more than just a few predictor to test. [5] proposed a penalized procedure via the LASSO penalty; [6] presented a unified variable selection method via the adaptive LASSO. But these methods are for the varying coefficient models. [7] established a model selection and semiparametric estimation method for additive quantile regression models by two-fold penalty. To our knowledge, the model selection and variable selection simultaneously with longitudinal data have not been investigated. We make several novel contributions: 1) We develop a new strategies for model selection and variable selection in additive model with longitudinal data; 2) We develop theoretical properties for our procedure.

In the next section, we will propose the two-fold SCAD penalization procedure based on QIF and computational algorithm; furthermore we present its theoretical properties. In particular, we show that the procedure can select the true model with probability approaching one, and show that newly proposed method estimates the non-zero function components in the model with the same optimal mean square convergence rate as the standard spline estimators. Simulation studies and an application of proposed methods in a real data example are included in Sections 3 and 4, respectively. Technical lemmas and proofs are given in **Appendix**.

2. Methodology and Asymptotic Properties

2.1. Additive Models with Two Fold Penalized Splines

Consider a longitudinal study with n subjects and n_i observations over time for the ith subject $(i=1,\cdots,n)$ for a total of $N=\sum_{i=1}^{n} n_i$ observation. Each observation consists of a response variable Y_{ij} and a covariate vector $X_{ij}\in R^d$ taken from the ith subject at time t_{ij}. We assume that the full data set

$$\left\{\left(X_{ij},Y_{ij}\right), i=1,\cdots,n,\ j=1,\cdots,n_i\right\}$$

is observed and can be modelled as

$$Y_{ij}=\mu+\sum_{l=1}^{d} m_l\left(X_{ij}^{(l)}\right)+\varepsilon_{ij},\quad i=1,\cdots,n,\ j=1,\cdots,n_i, \tag{2}$$

where ε_{ij} is random error with $E\left(\varepsilon_{ij}\middle|X_{ij}\right)=0$ and σ_ε^2.

At the start of the analysis, we do not know which component functions in model (1.1) are linear or actually zero. We adopt the centered B-spline basis, where $\boldsymbol{B}(\boldsymbol{X})=\left\{B_{s,l}\left(x_l\right):1\le l\le d,1\le s\le J\right\}^{\mathrm{T}}$ is a basis system $B_{s,l}\left(x_l\right)=\sqrt{K}\left[b_{s+1,l}\left(x_l\right)-\left\{E\left(b_{s+1,l}\right)/E\left(b_{1,l}\right)\right\}b_{1,l(x_l)}\right]$ and $(x)=\left(x_l\right)_{l=1}^{d}$. Equally-spaced knots are used in this article for simplicity of proof. Other regular knot sequences can also be used, with similar asymptotic results. Suppose that $m_l(\cdot)$ can be approximated well by a spline function, so that

$$m_l\left(x^{(l)}\right) \approx m_l^{sp}\left(x^{(l)}\right) = \sum_{s=1}^{J} \beta_{sl} B_{s,l}\left(x^{(l)}\right). \tag{3}$$

To simplify notation, we first assume equal cluster size $n_i = m < \infty$, and let $\beta_l = \left(\beta_{1l}, \cdots, \beta_{Jl}\right)^{\mathrm{T}}$, $\beta = \left\{\beta_{00}, \beta_1^{\mathrm{T}}, \cdots, \beta_d^{\mathrm{T}}\right\}_{Jd+1}^{\mathrm{T}}$ be the collection of the coefficients in (2.3), and $\mu = \beta_{00}$, denote $\boldsymbol{B}_{ij}^{(l)} = \left\{B_{1,l}\left(X_{ij}^{(l)}\right), \cdots, B_{J,l}\left(X_{ij}^{(l)}\right)\right\}_{J\times 1}^{\mathrm{T}}$ and $\boldsymbol{B}_{ij} = \left\{1, \boldsymbol{B}_{ij}^{(1)\mathrm{T}}, \cdots, \boldsymbol{B}_{ij}^{(d)\mathrm{T}}\right\}_{(Jd+1)\times 1}^{\mathrm{T}}$, then we have an approximation $\mu + m\left(\boldsymbol{X}_{ij}\right) = \boldsymbol{B}_{ij}^{\mathrm{T}}\beta$. We can also write the approximation of (2.1) in matrix notation as

$$\boldsymbol{Y}_i = \boldsymbol{B}_i^{\mathrm{T}}\beta + \varepsilon_i, \tag{4}$$

where $\boldsymbol{B}_i = \left\{\boldsymbol{B}_{i,1}, \cdots, \boldsymbol{B}_{im}\right\}_{n_i\times(Jd+1)}^{\mathrm{T}}$, $\boldsymbol{Y}_i = \left\{Y_{i1}, Y_{i2}, \cdots, Y_{im}\right\}^{\mathrm{T}}$ and $\varepsilon_i = \left\{\varepsilon_{i1}, \varepsilon_{i2}, \cdots, \varepsilon_{im}\right\}^{\mathrm{T}}$. [8] introduced the QIF that approximates the inverse of $\boldsymbol{R}$ by a linear combination of some basis matrixes, *i.e.*

$$\boldsymbol{R}^{-1} \approx a_0\boldsymbol{I} + a_1\boldsymbol{M}_1 + \cdots + a_K\boldsymbol{M}_K,$$

where $\boldsymbol{I}$ is the identity and $\boldsymbol{M}_i$ are known symmetric basis matrices and $a_0, a_1, \cdots, a_K$ are unknown constants. The advantage of the QIF approach is that it does not require the estimation of the linear coefficients a_i's associated with the working correlation matrix, which are treated as nuisance parameters here.

$$\boldsymbol{G}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\beta) = \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1}\left\{\boldsymbol{Y}_i - \boldsymbol{B}_i\beta\right\} \\ \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{A}_i^{-1/2}\left\{\boldsymbol{Y}_i - \boldsymbol{B}_i\beta\right\} \\ \vdots \\ \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_K\boldsymbol{A}_i^{-1/2}\left\{\boldsymbol{Y}_i - \boldsymbol{B}_i\beta\right\} \end{pmatrix}. \tag{5}$$

The vector $\boldsymbol{G}_n(\beta)$ contains more estimating equations than parameters, but these estimating equations can be combined optimally using the generalised method of the moment. So according to [8], the QIF approach estimates β by setting G_n as close to zero as possible, in the sense of minimizing the quadratic inference function $\boldsymbol{Q}_n(\beta)$.

$$\boldsymbol{Q}_n(\beta) = n\boldsymbol{G}_n^{\mathrm{T}}(\beta)\boldsymbol{C}_n^{-1}(\beta)\boldsymbol{G}_n(\beta), \tag{6}$$

where

$$\boldsymbol{C}_n^{-1}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\beta)\boldsymbol{g}_i^{T}(\beta).$$

Our main goal is to find both zero components (*i.e.*, $m_j \equiv 0$) and linear compoents (*i.e.*, m_j is a linear function). The former can be achieved by shrinking $\left\|m_j^{sp}\right\|$ to zero. For the latter, we want to shrink the second derivative $\left\|m_j^{sp(")}\right\|$ to zero instead. This suggests the following minimization problem

$$\hat{\beta} = \arg\min_{\beta} \boldsymbol{Q}_n(\beta) + n\sum_{l=l}^{d} p_{\lambda_1}\left(\left\|m_l^{sp}\right\|\right) + n\sum_{l=1}^{d} p_{\lambda_2}\left(\left\|m_l^{sp(")}\right\|\right), \tag{7}$$

where $p_{\lambda_1}(\cdot)$ and $p_{\lambda_2}(\cdot)$ are two penalties used to find zero and linear coefficients respectively, with two regularization parameters λ_1 and λ_2, and $m_l^{sp} = \beta_l^{\mathrm{T}}\boldsymbol{B}^{(l)}$, $\boldsymbol{B}^{(l)} = \left\{B_{1,l}, B_{2,l}, \cdots, B_{J,l}\right\}^{\mathrm{T}}$. Note that since $\left\|m_l^{sp}\right\|^2 = \left\|\hat{\beta}_l^{\mathrm{T}}\boldsymbol{B}^{(l)}\right\|^2 = \int\left(\sum_k \beta_{kl}B_{kl}(x)\right)\left(\sum_{k'}\beta_{k'l}B_{k'l}(x)\right)\mathrm{d}x$ and

$$\left\|m_l^{sp(")}\right\|^2 = \left\|\hat{\beta}_l^{\mathrm{T}}\boldsymbol{B}^{(l)}\right\|^2 = \int\left(\sum_k \beta_{kl}B_{kl}^{(")}(x)\right)\left(\sum_{k'}\beta_{k'l}B_{k'l}^{(")}(x)\right)\mathrm{d}x,$$

$\left\|m_l^{sp}\right\|$ and $\left\|m_l^{sp(")}\right\|$ can be equivalently written as $\|\beta_l\|_{\boldsymbol{D}_l} = \sqrt{\beta_l^{\mathrm{T}}\boldsymbol{D}_l\beta_l}$ and $\|\beta_l\|_{\boldsymbol{E}_l} = \sqrt{\beta_l^{\mathrm{T}}\boldsymbol{E}_l\beta_l}$ respectively,

with (k,k') entry of $\boldsymbol{D}_l$ being $\int_0^1 B_{kl}(x)B_{k'l}(x)\mathrm{d}x$.

2.2. Asymptotic Properties

To study the rate of convergence for $\hat{\mu}$ and $\hat{\beta}$, we first introduce some notations and present regularity conditions. We assume equal cluster sizes $(n_i = m < \infty)$, and $(Y_i, X_i), i = 1,\cdots,n$ are i.i.d. from (Y, X) with $\boldsymbol{Y}_i = (Y_{i1},\cdots,Y_{im})^{\mathrm{T}}$, and $\boldsymbol{X}_i = (X_{i1}^{\mathrm{T}},\cdots,X_{im}^{T})^{\mathrm{T}}$. For convenience, we assume that $m_j(\cdot)$ is truly nonparametric for $1 \le j \le d_1$, is linear for $d_1+1 \le j \le d_1+d_2 = s$, and is zero for $s+1 \le j \le d$. The asymptotic result still hold for data with unequal cluster sizes m_i using a cluster-specific transformation as discuss in [4]. For any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes the modulus of the largest singular value of $\boldsymbol{A}$. To prove the theoretical arguments, we need the following assumptions:

(A1) The covariates $\boldsymbol{X}_i = \{X_{i1}^{\mathrm{T}},\cdots,X_{im}^{\mathrm{T}}\}^{\mathrm{T}}$ are compactly supported, and without loss of generality, we assume that each X_{ij}^{T} has support $\chi = [0,1]^d$. The density of X_{ij}^{T}, denoted by $f_j(\boldsymbol{x})$, is absolutely continuous and there exist constants C_1 and C_2 such that $0 < C_1 \le \min_{x\in\chi} f_j(\boldsymbol{x}) \le C_2 < \infty$ for all $j = 1,\cdots,m$.

(A2) Let $e = Y - \mu - m(\boldsymbol{X})$. Then $\tilde{\Sigma} = Eee^{\mathrm{T}}$ is positive definite and for some $\delta > 0$, $E\|e\|^{2+\delta} < +\infty$.

(A3) For each $l = 1,\cdots,d$, $m_l(\cdot)$ has r continuous derivatives for some $r \ge 2$.

(A4) $$\boldsymbol{G}_n(\beta) = \lim_{n\to\infty}\frac{1}{n}E\begin{pmatrix}\boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_1\boldsymbol{A}_i^{-1/2}\boldsymbol{B}_i\\ \vdots \\ \boldsymbol{B}_i^{\mathrm{T}}\boldsymbol{A}_i^{-1/2}\boldsymbol{M}_K\boldsymbol{A}_i^{-1/2}\boldsymbol{B}_i\end{pmatrix} = J_0. \tag{8}$$

(A5) Let $\boldsymbol{M} = (\boldsymbol{M}_0^{\mathrm{T}},\cdots,\boldsymbol{M}_k^{\mathrm{T}})^{\mathrm{T}}$. Assume the modular of the singular value of $\boldsymbol{M}$ is bounded away from 0 and infinity.

(A6) The matrix $\boldsymbol{A}$ defined in Theorem 3 is positive definite.

Theorem 1. Suppose that the regularity conditions A1-A5 hold and the number of knots $K = O_p(n^{1/(2r+1)})$, $\lambda_1, \lambda_2 \to 0$. Then there exists a local minimizer of (2.7) such that

$$|\hat{\mu} - \mu_0| = O_p(n^{-r/(2r+1)}),$$

$$\max_{1\le l\le d}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\{\hat{m}_l(x) - m_l(x)\}^2 = O_p(n^{-2r/(2r+1)}).$$

For $m_i = m = 1$, it reduces to a special case where the responses are i.i.d. The rate of convergence given here is the same optimal rate as that obtain for polynomial spline regression for independent data [9] [10]. The main advantage of the QIF approach is that it incorporates within-cluster correlation by optimally combing estimating equations without estimating the correlation parameters. the estimator of two fold penalized QIF achieve the same rate of convergence as un-penalized estimator. Furthermore, we prove that the penalized estimators $\{\hat{\beta}_l\}_{l=1}^{d}$ in Theorem 1 possess the sparsity property, $\hat{m}_l = 0$ almost surely for $l = s+1,\cdots,d$. The sparsity property ensures that the proposed model selection is consistent, that is, it selects the correct variables with probability tending to 1 as the sample size goes to infinity.

Theorem 2. Under the same assumptions of Theorem 1, and if the tuning parameter $n^{r/(2r+1)}\min\{\lambda_1,\lambda_2\} \to \infty$. Then with probability approaching 1.

a) $\hat{m}_j \equiv 0, s+1 \le j \le d$

b) $\hat{m}_j$ is a linear function for $d_1+1 \le j \le s$

Theorem 2 also implies that above additive model selection possesses the consistency property. The results in Theorems 2 are similar to semiparametric estimation of additive quantile regression model in [7]. However, the theoretical proof is very different from the penalized quantile loss function due to the two fold penalty and longitudinal data.

Finally, in the same spirit of that [11], we come to the question of whether the SIC can identify the true model in our setting.

Theorem 3. Suppose that the regularity conditions A1-A5 hold and the number of knots $K = O_p\left(n^{1/(2r+1)}\right)$ as assumed in Theorem 1, The parameters $\hat{\lambda}_1$ and $\hat{\lambda}_2$ selected by SIC can select the true model with probability tending to 1.

3. Simulation Study

In this section, we conducted Monte Carlo studies for the following longitudinal data and additive model. the continuous responses $\{Y_{ij}\}$ are generated from

$$Y_{ij} = \sum_{l=1}^{d} m_l\left(X_{ij}^{(l)}\right) + \varepsilon_{ij}, \quad 1,\cdots,n, \quad j = 1,\cdots,5 \tag{9}$$

where $d = 10$ and the number of clusters $n = 100, 250, 500$. The additive functions are $m_1(x) = 5\sin(2\pi x)/(2-\sin(2\pi x)), m_2(x) = 8(x-0.5)^2, m_3(x) = 2x, m_4(x) = x, m_5(x) = -x$. Thus the last 5 variables in this model are null variables and do not contribute the model. The covariates $\boldsymbol{X}_{ij} = \left(X_{ij}^{(1)},\cdots,X_{ij}^{(10)}\right)^{\mathrm{T}}$ are generated independently from uniform. The error $\varepsilon_i = (\varepsilon_{i1},\cdots,\varepsilon_{i6})^{\mathrm{T}}$ follows a multivariate normal distribution with mean 0, a common marginal variance $\sigma^2 = 1$, it has first-order autoregressive (AR-1) and an compound symmetry (CS) correlation (*i.e.* exchangeable correlation) structure with different within correlation coefficient, and consider $\rho = 0.8$ and $\rho = 0.3$ representing a strong and weak within correlation structure. The predictors $\boldsymbol{X}_{ij} = \left(X_{ij}^{(1)},\cdots,X_{ij}^{(10)}\right)^{\mathrm{T}}$ are generated by $X_{ij}^{(l)} = \Phi\left(Z_{ij}^{(l)}\right)$, $\boldsymbol{Z}_{ij} = \left(Z_{ij}^{(1)},\cdots,Z_{ij}^{(l)}\right) \sim N(0,\Sigma)$, $i = 1,\cdots,n, j = 1,\cdots,m$, where Φ is the standard normal c.d.f. and $\Sigma = (1-r)\boldsymbol{I}_{(d\times d)} + r\mathbf{1}_d\mathbf{1}_d^{\mathrm{T}}$. The parameter $r\ (0 \le r < 1)$ controls the correlation between $Z_{ij}^{(l)}, 1 \le (l) \le d$.

To illustrate the effect on estimation efficiency, we compare the penalized QIF approach in [4] (PQIF) and an Oracle model (ORACLE). here the full model consists of all ten variable, and oracle model only contains the first five relevant variables and we know it's a partial additive model. The oracle model is only available in simulation studies where the true information is known. In all simulation, the number of replications is 100 and the result are summarized in **Table 1** and **Table 2**. In **Table 1**, the model selection result for both our procedure

Table 1. The estimation results for our estimator (TFPQIF) and sparse additive estimator (PQIF) and ORACLE esitmator.

n	Correlation	Method	μ	m_1	m_2	m_3	m_4	m_5
100	CS	PQIF	0.32	0.42	0.3	0.29	0.31	0.26
		TFPQIF	0.3	0.46	0.28	0.25	0.23	0.22
		ORACLE	0.14	0.14	0.15	0.15	0.13	0.12
	AR(1)	PQIF	0.36	0.39	0.32	0.3	0.29	0.25
		TFPQIF	0.29	0.39	0.35	0.2	0.25	0.22
		ORACLE	0.13	0.15	0.22	0.14	0.12	0.1
250	CS	PQIF	0.25	0.29	0.25	0.24	0.19	0.15
		TFPQIF	0.22	0.31	0.26	0.14	0.16	0.15
		ORACLE	0.12	0.11	0.19	0.097	0.098	0.09
	AR(1)	PQIF	0.28	0.24	0.31	0.33	0.28	0.19
		TFPQIF	0.20	0.2	0.27	0.24	0.14	0.15
		ORACLE	0.1	0.11	0.13	0.21	0.1	0.096
500	CS	PQIF	0.15	0.14	0.25	0.23	0.2	0.17
		TFPQIF	0.15	0.3	0.26	0.11	0.12	0.1
		ORACLE	0.09	0.13	0.12	0.07	0.07	0.07
	AR(1)	PQIF	0.18	0.3	0.26	0.13	0.12	0.14
		TFPQIF	0.16	0.23	0.25	0.09	0.09	0.09
		ORACLE	0.08	0.13	0.12	0.077	0.081	0.07

Table 2. Model selction results for our estimator (TFPQIF) and sparse additive estimator (PQIF) and ORACLE esitmator.

λ_0	λ_1	CS				AR(1)			
		NCC	NNT	NLC	NLT	NCC	NNT	NLC	NLT
100	PQIF	5.96	2	0	0	5.83	2	0	0
	TFPQIF	2.64	2	2.58	2.36	2.52	2	2.63	2.46
	ORACLE	2	2	3	3	2	2	3	3
250	PQIF	5.63	2	0	0	5.45	2	0	0
	TFPQIF	2.34	2	2.66	2.65	2.41	2	2.59	2.5
	ORACLE	2	2	3	3	2	2	3	3
500	PQIF	5.35	2	0	0	5.2	2	0	0
	TFPQIF	2.04	2	2.93	2.93	2.1	2	2.89	2.86
	ORACLE	2	2	3	3	2	2	3	3

with the one penalty QIF when the error are Gaussian, and we also list the oracle model as a benchmark, the oracle model is only available in simulation studies where the true information is known in **Table 1**, in which the column labeled "NNC" presents the average number of nonparametric components selected, the column "NNT" depicts the average number of nonparametric components selected that are truly nonparametric (truly nonzero for one penalty QIF), "NLC" presents the average number of linear components, "NLT" depicts the average number of linear components selected that are truly linear.

In **Table 2**, we conduct some simulations to evaluate finite sample performance of the proposed method. Let $\hat{m}_k(\cdot)$ be the estimator of a nonparametric function $m_k(\cdot)$ and $\{u_s\}_{s=1}^{M}$ be the grid points, the performance of estimator $\hat{m}_j(\cdot)$ will be assessed by using the square root of average square errors(RASE), we compare the performance of above estimators. On the nonparametric coponents, the errors for estimators with a single penalty and our procedure are similar, and both are qualitatively close to those of the oracle estimator. However, for the parametric components, our estimator is obviously more efficient,leading to about 40% - 50% reduction in RASE.

$$RASE = \left\{ \frac{1}{M} \sum_{s=1}^{M} \sum_{k=1}^{d} \left[\hat{m}_k(u_s) - m_k(u_s) \right]^2 \right\}^{1/2}.$$

4. Real Data Analysis

In this subsection, we analyze data from the Multi-Center AIDS Cohort Study. The dataset contains the human immunodeficiency virus, HIV, status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. All individuals were scheduled to have their measurements made during semi-annual visits. Here $t_{ij}, i = 1, \cdots, n, j = 1, \cdots, m_i$ denotes the time length in years between seroconversion and the j-th measurement of the i-th individual after the infection. [12] analyzed the dataset using partial linear models. The primary interest was to describe the trend of the mean CD4 percentage depletion over time and to evaluate the effects of cigarette smoking, pre-HIV infection CD4 percentage, and age at infection on the mean CD4 cell percentage after the infection.

In our analysis, the response variable is the CD4 cell percentage of a subject at distinct time points after HIV infection. We take four covariates for this study: X_1, the CD4 cell percentage level before HIV infection; and X_2, age at HIV infection; X_3 the individual's smoking status, which takes binary values 1 or 0, according to whether a individual is a smoker or nonsmoker; T_{ij} denote $i = 1, \cdots, n, j = 1, \cdots, m$, denotes the time length in years between seroconversion and the j-th measurement of the i-th individual after the infection. We construct the following additive model;

$$Y_{ij} = \mu + m_1\left(X_{(1)}^{ij}\right) + m_2\left(X_{ij}^{(2)}\right) + m_3\left(T_{ij}\right) + \beta_0 X_{ij}^{(3)} + \epsilon_{ij}.$$

the partially linear additive models instead of additive model because of the binaray variable $X^{(3)}$, but we not select the linear component. using our procedure, we wang to ensure which is linear component and which is zero in the non-parametirc function. For implement our procedure, linear transformation be used to the variable $X^{(1)}, X^{(2)}, T$. The result of our procedure select the m_1 is zero function and select the m_2 is a linear function, m_3 is a non-parametric. As shown in **Figure 1**, we see that the mean baseline CD4 percentage of the population

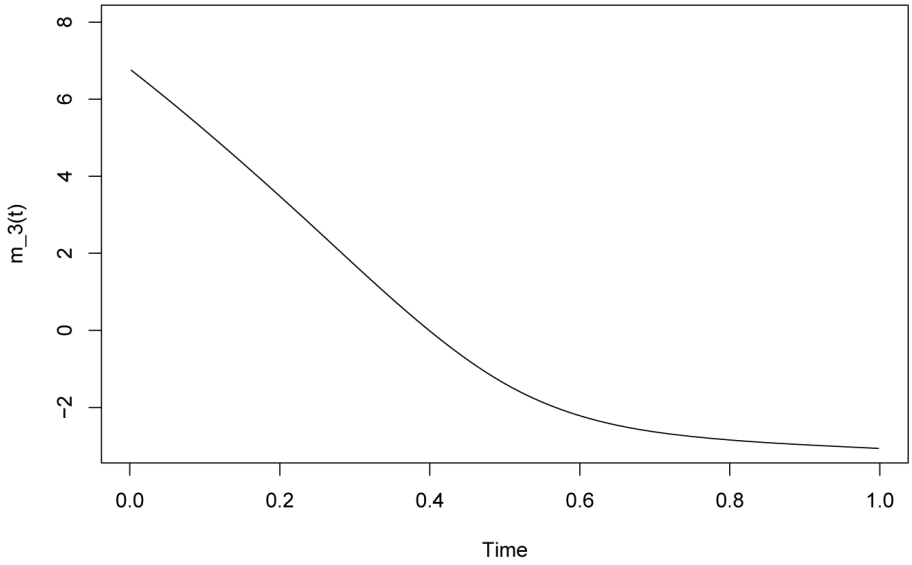

Figure 1. The estimator of $m_3(x)$.

depletes rather quickly at the beginning of HIV infection, but the rate of depletion appears to be slowing down at four years after the infection. This result is the same as before [13].

5. Concluding Remark

In summary, we present a two-fold penalty variable selection procedure in this paper, which can select linear component and significant covariate and estimate unknown coefficient function simultaneously. The simulation study shows that the proposed model selection method is consistent with both variable selection and linear components selection. Besides, being theoretically justified, the proposed method is easy to understand and straightforward to implement. Further study of the problem is how to use the multi-fold penalty to solve the model selection and variable selection in generalized additive partial linear models with longitudinal data.

Acknowledgements

Liugen Xue's research was supported by the National Nature Science Foundation of China (11171012), the Science and Technology Project of the Faculty Adviser of Excellent PhD Degree Thesis of Beijing (20111000503) and the Beijing Municipal Education Commission Foundation (KM201110005029).

References

[1] Hastie, T.J. and Tibshirani, R.J. (1990) Generalized Additive Models. Chapman and Hall, London.

[2] Berhane, K. and Tibshirani, R.J. (1998) Generalized Additive Models for Longitudinal Data. *The Canadian Journal of Statistics*, **26**, 517-535. http://dx.doi.org/10.2307/3315715

[3] Martinussen, T. and Scheike, T.H. (1999) A Semiparametric Additive Regression Model for Longitudinal Data. *Biometrika*, **86**, 691-702. http://dx.doi.org/10.1093/biomet/86.3.691

[4] Xue, L. (2010) Consistent Model Selection for Marginal Generalized Additive Model for Correlated Data. *Journal of the American Statistical Association*, **105**, 1518-1530. http://dx.doi.org/10.1198/jasa.2010.tm10128

[5] Hu, T. and Xia, Y.C. (2012) Adaptive Semi-Varying Coefficient Model Selection. *Statistica Sinica*, **22**, 575-599. http://dx.doi.org/10.5705/ss.2010.105

[6] Tang, Y.L., Wang, H.X., Zhu, Z.Y. and Song, X.Y. (2012) A Unified Variable Selection Approach for Varying Coefficient Models. *Statistica Sinica*, **22**, 601-628. http://dx.doi.org/10.5705/ss.2010.121

[7] Lian, H. (2012) Shrinkage Estimation for Identification of Linear Components in Additive Models. *Statistics and Probability Letters*, **82**, 225-231. http://dx.doi.org/10.1016/j.spl.2011.10.009

[8] Qu, A., Lindsay, B.G. and Li, B. (2000) Improving Generalised Estimating Equations Using Quadratic Inference Func-

tions. *Biometrika*, **87**, 823-836. http://dx.doi.org/10.1093/biomet/87.4.823

[9] Huang, J.Z. (1998) Projection Estimation in Multiple Regression with Application to Functional ANOVA Models. *The Annals of Statistics*, **26**, 242-272. http://dx.doi.org/10.1214/aos/1030563984

[10] Xue, L. (2009) A Root-N Consistent Backfitting Estimator for Semiparametric Additive Modeling. *Statistica Sinica*, **19**, 1281-1296.

[11] Wang, H., Li, R. and Tsai, C.L. (2007) Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika*, **94**, 553-568. http://dx.doi.org/10.1093/biomet/asm053

[12] Xue, L.G. and Zhu, L.X. (2007) Empirical Likehood for a Varying Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, **102**, 642-654.

[13] Wu, C.O., Chiang, C.T. and Hoover, D.R. (2010) Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, **93**, 1388-1402.

[14] De Boor, C. (2001) A Practical Guide to Splines. Springer, New York.

Appendix: Proofs of Theorems

For convenience and simplicity, let C denote a positive constant that may be different at each appearance throughout this paper. Before we prove our main theorems, we list some regularity conditions that are used in this paper.

Lemma 1. Under the conditions (A1)-(A6), minimizing the no penalty QIF $\tilde{\beta} = \arg\min_{\beta} \boldsymbol{Q}_n(\beta)$. Then $|\tilde{\mu} - \mu_0| = O\left(n^{-\frac{2r}{2r+1}}\right)$, $\max_{1\le l\le d} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\tilde{m}_l(x_{ij}) - m_l(x_{ij})\right\}^2 = O_p\left(n^{-2r/(2r+1)}\right)$.

Proof: According to [14], for each $l = 1,\cdots,d$, we can get $m = \mu + \sum_{l=1}^{d} m_l$ satisfying the condition (4). There exists a constant $C > 0$ and a spline function $\tilde{m} \in \mathcal{G}_n$, such that $\|\tilde{m} - m\|_{\infty} \le CK^{-r}$. Using the triangular in equality $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\tilde{m}_l(x_{ij}) - m_l(x_{ij})\right\}^2 \le \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\boldsymbol{B}_l(x_{ij})\left(\tilde{\beta}_l - \beta_l^0\right)\right\}^2 + O\left(K^{-2r}\right)$. Therefore, it is sufficient to show that $\left\|\boldsymbol{B}_l\left(\tilde{\beta}_l - \beta_l^0\right)\right\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\left\{\boldsymbol{B}_l(x_{ij})\left(\tilde{\beta}_l - \beta_l^0\right)\right\}^2 = O_p\left(n^{-1}K\right)$. According to [8] entail that for any $\epsilon > 0$. exists sufficiently large $C > 0$. such that $n \to \infty$ $p\left\{\inf_{\|\boldsymbol{B}(\tilde{\beta}-\beta^0)\|_n \le C(K/n)^{1/2}} \boldsymbol{Q}_n(\beta) > \boldsymbol{Q}_n(\beta_0)\right\} > 1-\epsilon$, therefore $p\left\{\left\|\boldsymbol{B}\left(\tilde{\beta} - \beta^0\right)\right\|_n = O_p\left(\left(\frac{K}{n}\right)^{1/2}\right)\right\}$. Furthermore, for each $l = 1,\cdots,d$. There exists a constant $C > 0$. such that $\left\|\boldsymbol{B}\left(\tilde{\beta}_l - \beta_l^0\right)\right\|_n^2 \le C\left\|\boldsymbol{B}\left(\tilde{\beta} - \beta^0\right)\right\|_n^2 = O_p\left(\left(\frac{K}{n}\right)^{1/2}\right)$.

Proof of Theorem 1. Let

$$\boldsymbol{L}_n(\beta) = \boldsymbol{Q}_n(\beta) + n\sum_{j=1}^{d} p_{\lambda_1}\left(\left\|\tilde{m}_j\right\|\right) + n\sum_{j=1}^{d} p_{\lambda_2}\left(\left\|\tilde{m}_j''\right\|\right)$$

be the object function in (2.7), where $\|\beta_j\|_{\boldsymbol{D}_j} = \sqrt{\beta_j^{\mathrm{T}}\boldsymbol{D}_j\beta_j}$ and $\|\beta_j\|_{\boldsymbol{E}_j} = \sqrt{\beta_j^{\mathrm{T}}\boldsymbol{E}_j\beta_j}$, as a special case of no penalty QIF. Let $\tilde{\beta} = \arg\min_{\beta} \boldsymbol{Q}_n(\beta)$, and $\tilde{m}_j = \boldsymbol{B}^{\mathrm{T}}\tilde{\beta}_j$, well known result is $\left\|\boldsymbol{B}^{\mathrm{T}}\left(\tilde{\beta} - \beta_0\right)\right\|_n = O_p\left(\sqrt{K/n}\right)$, we want to show that for large n and any $\varepsilon > 0$, there exist a constant C large enough such that

$$p\left\{\inf_{\beta:\left\|\boldsymbol{B}^{\mathrm{T}}(\tilde{\beta}-\beta_0)\right\| = C\left(nK^{-1}\right)^{-1/2}} \boldsymbol{L}_n(\beta) > \boldsymbol{L}_n\left(\tilde{\beta}\right)\right\} \ge 1-\epsilon. \tag{A1}$$

As a result, this implies that $\boldsymbol{L}_n(\cdot)$ has a local minimum in the ball $\left\{\beta : \left\|\boldsymbol{B}^{\mathrm{T}}\left(\beta - \tilde{\beta}\right)\right\|\right\}$. Thus, $\left\|\boldsymbol{B}^{\mathrm{T}}\left(\beta - \tilde{\beta}\right)\right\|_n = O_p\left(1/\sqrt{nK^{-1}}\right)$. Further, the triangular inequality gives $\left\|\boldsymbol{B}^{\mathrm{T}}\hat{\beta} - m\right\|_n = O_p\left(\left(nK^{-1}\right)^{-1/2} + K^{-r}\right)$. To show (A1), For convenience, we assume that m_j is truly nonparametric for $1 \le j \le d_1$ is linear for $d_1 + 1 \le j \le s = d_1 + d_2$ and zero for $s+1 \le j \le d$.

$$\begin{aligned}\boldsymbol{L}_n(\beta) - \boldsymbol{L}_n\left(\tilde{\beta}_j\right) \ge \boldsymbol{Q}_n(\beta) - \boldsymbol{Q}_n\left(\tilde{\beta}_j\right) &+ \sum_{j=1}^{s} n\left\{p_{\lambda_1}\left(\|\beta_j\|_{\boldsymbol{D}_j}\right) - p_{\lambda_1}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{D}_j}\right)\right\} \\ &+ \sum_{j=1}^{d_1} n\left\{p_{\lambda_2}\left(\|\beta_j\|_{\boldsymbol{E}_j}\right) - p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\right\}.\end{aligned} \tag{A2}$$

Since $\left\|\hat{\beta}_j - \beta_{0j}\right\|_n$. Since $\lambda_1, \lambda_2 = o(1)$. We have $p\left(p_{\lambda_1}\left(\|\beta_j\|_{\boldsymbol{D}_j}\right) = p_{\lambda_1}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{D}_j}\right)\right) \to 1$. If $i \le s = d_1 + d_2$,

similarly, $p\left(p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)=p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\right)\to 1$. If $l\le d_1$. These facts imply that $n\sum_{j=1}^{d}p_{\lambda_1}\left(\left\|\beta_j\right\|_{\boldsymbol{D}_j}\right)-n\sum_{j=1}^{d}p_{\lambda_1}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{D}_j}\right)\ge 0$ and $n\sum_{j=1}^{d}p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)-n\sum_{j=1}^{d}p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\ge 0$ with probability tending to 1. If $\lambda_1,\lambda_2=o(1)$, $\left\|\beta_j\right\|_{\boldsymbol{D}_j}\ge C\lambda_{\max}\left(D_j\right)$, $\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\ge C\lambda_{\max}\left(E_j\right)$, for $i=1,\cdots,d$. Therefore, when n is large enough,

$$\sum_{j=1}^{d}n\left\{p_{\lambda_1}\left(\left\|\beta_j\right\|_{\boldsymbol{D}_j}\right)-p_{\lambda_1}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{D}_j}\right)\right\}=C_1nK^{-1/2}\lambda_1\sum_j\left\|\beta_j-\tilde{\beta}_j\right\|_n\to 0$$

$$\sum_{j=1}^{d}n\left\{p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)-p_{\lambda_2}\left(\left\|\tilde{\beta}_j\right\|_{\boldsymbol{E}_j}\right)\right\}=C_2nK^{-1/2}\lambda_2\sum_j\left\|\beta_j-\tilde{\beta}_j\right\|_n\to 0$$

By the definition of SCAD penalty function, removing the regularizing terms in (A2)

$$\boldsymbol{Q}_n(\beta)-\boldsymbol{Q}_n\left(\tilde{\beta}_j\right)=\left(\beta-\tilde{\beta}\right)^{\mathrm{T}}\dot{\boldsymbol{Q}}_n\left(\tilde{\beta}_j\right)+\frac{1}{2}\left(\beta-\tilde{\beta}\right)^{\mathrm{T}}\ddot{\boldsymbol{Q}}_n\left(\tilde{\beta}\right)\left(\beta-\tilde{\beta}\right)\left\{1+o_p(1)\right\}\tag{A3}$$

with $\dot{\boldsymbol{Q}}_n$ and $\ddot{\boldsymbol{Q}}_n$ being the gradient vector and hessian matrix $\boldsymbol{Q}_n$, respectively. Following [8], and Lemma A1 in supplement, for any β with $\left\|\boldsymbol{B}^{\mathrm{T}}\left(\beta-\tilde{\beta}\right)\right\|_n=C\left(nK^{-1}\right)^{-1/2}$, one has

$$n\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\dot{\boldsymbol{Q}}_n\left(\tilde{\beta}_j\right)=n\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\dot{\boldsymbol{G}}_n^{\mathrm{T}}\left(\tilde{\beta}_j\right)\boldsymbol{C}_n^{-1}\left(\tilde{\beta}_j\right)\boldsymbol{G}_n\left(\tilde{\beta}_j\right)\left\{1+o_p(1)\right\}=O(K)$$

and

$$\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\ddot{\boldsymbol{Q}}_n\left(\tilde{\beta}_j\right)\left(\beta-\tilde{\beta}_j\right)=n\left(\beta-\tilde{\beta}_j\right)^{\mathrm{T}}\dot{\boldsymbol{G}}_n^{\mathrm{T}}\left(\tilde{\beta}_j\right)\boldsymbol{C}_n^{-1}\left(\tilde{\beta}_j\right)\dot{\boldsymbol{G}}_n\left(\tilde{\beta}_j\right)\left(\beta-\tilde{\beta}_j\right)\left\{1+o_p(1)\right\}=O(K)$$

where $\dot{\boldsymbol{G}}_n$ is the first order derivative of $\boldsymbol{G}_n$. Therefore, by choosing C large enough, the second term on (A3) dominates its first term. therefore (A1) holds when C and n are large enough. This completes the proof of Theorem 1. □

Proof of Theorem 2. We only show part (b), as an illustration and part (a) is similar. Suppose for some $d_1<j\le s$, $B_j^{\mathrm{T}}\hat{\beta}_j$ does not represent a linear function. Define $\hat{\beta}_j^{\star}$ to be the same as $\hat{\beta}$ except that $\hat{\beta}_j$ is replaced by its projection onto the subspace $\{\beta_j:B_j^{\mathrm{T}}\beta_j$ represents a linear function$\}$, we have

$$\boldsymbol{L}_n\left(\hat{\beta}\right)-\boldsymbol{L}_n\left(\tilde{\beta}_j\right)=\boldsymbol{Q}_n\left(\hat{\beta}\right)-\boldsymbol{Q}_n\left(\hat{\beta}_j^{\star}\right)+\sum_j n\left\{p_{\lambda_1}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{D}_j}\right)-p_{\lambda_1}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{D}_j}\right)\right\}$$
$$+\sum_j n\left\{p_{\lambda_2}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}\right)-p_{\lambda_2}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{E}_j}\right)\right\}.$$

As in the proof of Theorem 1, we have $P\left(p_{\lambda_1}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{D}_j}\right)-p_{\lambda_1}\left(\left\|\hat{\beta}_j^{\star}\right\|_{\boldsymbol{D}_j}\right)\right)\to 1$ and thus with probability approaching 1

$$0\ge\boldsymbol{L}_n\left(\hat{\beta}\right)-\boldsymbol{L}_n\left(\tilde{\beta}_j\right)=\boldsymbol{Q}_n\left(\hat{\beta}\right)-\boldsymbol{Q}_n\left(\tilde{\beta}_j\right)+n\sum_j p_{\lambda_2}\left(\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}\right).\tag{A4}$$

$\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}=\sqrt{\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}=\sqrt{\left(\hat{\beta}_j-\beta_{0j}\right)^{\mathrm{T}}\boldsymbol{E}_j\left(\hat{\beta}_j-\beta_{0j}\right)}=O_p\left(\left(\frac{K}{n}\right)^{1/2}\right)=o(\lambda_2)$. $p_{\lambda_2}\left(\left\|\beta_j\right\|_{\boldsymbol{E}_j}\right)=\lambda_2\left\|\hat{\beta}_j\right\|_{\boldsymbol{E}_j}$, with probability tending to 1. by the definition of SCAD penalty. $\left\|\hat{\beta}_j-\hat{\beta}_j^{\star}\right\|=O_p\left(\sqrt{K\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}\right)$, $np_{\lambda_2}\left(\sqrt{\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}\right)=n\lambda_2\sqrt{K\hat{\beta}_j^{\mathrm{T}}\boldsymbol{E}_j\hat{\beta}_j}$. Therefore, similar to the proof of Theorem 1, by choosing C large enough, the second term on the right had side of (A4) dominates its first term. □

Proof of Theorem 3. For any regularization parameters $\lambda = (\lambda_1, \lambda_2)$, we denote the estimator of two fold penalty $\hat{\beta}_\lambda$, and denote by $\hat{\beta}$ the minimizer when the optimal sequence of regularization parameters is chosen. There are four separate cases to consider

CASE 1: $B_j \beta_{\lambda j}$ represents a linear component for some $j \le d_1$. Similar to the proof of Theorems 1 and 2, we have

$$\boldsymbol{Q}_n(\hat{\beta}) - \boldsymbol{Q}_n(\hat{\beta}_\lambda) = (\hat{\beta} - \hat{\beta}_\lambda)^{\mathrm{T}} \dot{\boldsymbol{Q}}_n(\hat{\beta}_\lambda) + \frac{1}{2}(\hat{\beta} - \hat{\beta}_\lambda)^{\mathrm{T}} \ddot{\boldsymbol{Q}}_n(\hat{\beta})(\hat{\beta} - \hat{\beta}_\lambda)\{1 + o_p(1)\}.$$

Since true m_j not linear and $\hat{\beta}_j$ is consistent in model selection, $\dfrac{\|\hat{\beta} - \hat{\beta}_\lambda\|}{K}$ is bounded away form zero, thus $\log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta})\right\} - \log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda)\right\} > (C_1 K + C_2)\dfrac{\log(n)}{2n}$, for any $0 \le C_1, C_2 \le d$, with probability tending to 1 and the SIC cannot select such λ.

CASE 2: $\hat{\beta}_{\lambda j}$ is zero for some $1 \le j \le s$. The proof is very similar with CASE 1 and therefore omitted.

CASE 3: $\boldsymbol{B}_j \hat{\beta}_j$ represents a nonlinear component for some $d_1 < j \le s$. Here when considering CASE 3, we implicity exclude all previous cases that no underfitting cases. $\tilde{\beta}$ is the estimator of minimizing the no penalty QIF (2.6) $\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda) - \frac{1}{n}\boldsymbol{Q}_n(\tilde{\beta}) \ge -\left|O\left(\frac{K}{n}\right)\right|$. Thus $\log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda)\right\} - \log\left\{\frac{1}{n}\boldsymbol{Q}_n(\tilde{\beta}) \ge -\left|O\left(\frac{K}{n}\right)\right|\right\}$ and

$\log\left\{\frac{1}{n}\boldsymbol{Q}_n(\hat{\beta}_\lambda)\right\} + \dfrac{K\log(n)}{2n} \ge \log\left\{\frac{1}{n}\boldsymbol{Q}_n(\tilde{\beta})\right\} + \dfrac{d\log(n)}{2n}$ with probability tending to 1. □

CASE 4: $\hat{\beta}_{\lambda j}$ is nonzero for $j \ge s$. The case is similar to case 3. Thus the proof is omitted.