Hacettepe Journal of Mathematics and Statistics
Volume 47 (4) (2018), 1041 – 1060

Identification and estimation for generalized varying coefficient partially linear models

Mingqiu Wang*, Xiuli Wang† and Muhammad Amin‡ §

Abstract

The *generalized varying coefficient partially linear model* (GVCPLM) enjoys the flexibility of the generalized varying coefficient model and the parsimony and interpretability of the generalized linear model. Statistical inference of GVCPLM is restricted with a condition that the components of varying and constant coefficients are known in advance. Alternatively, the current study is focused on the structure's identification of varying and constant coefficient for GVCPLM and it is based on the spline basis approximation and the group SCAD. This is proved that the proposed method can consistently determine the structure of the GVCPLM under certain conditions, which means that it can accurately choose the varying and constant coefficients precisely. Simulation studies and a real data application are conducted to assess the infinite sample performance of the proposed method.

Keywords: Group variable selection, Group SCAD, Selection consistency, Structure identification.

Mathematics Subject Classification (2010): 62J07, 62G05, 62G08.

Received : 01.04.2016 *Accepted* : 22.09.2016 *Doi* : 10.15672/HJMS.201614821897

*School of Statistics, Qufu Normal University, Qufu, Shandong 273165, China, Email: wmqwxl@hotmail.com

†School of Statistics, Qufu Normal University, Qufu, Shandong 273165, China, Email: wang_qstj@hotmail.com

‡Nuclear Institute for Food and Agriculture (NIFA), 446, Peshawar, Pakistan, Email: aminkanju@gmail.com

§Corresponding Author.

1. Introduction

Semiparametric regression models have received great attention in many contemporary statistical studies because it combines the flexibility of nonparametric regression and the parsimony and interpretability of parametric regression. Considering a general specification of semiparametric models, the generalized varying coefficient partially linear model (GVCPLM) is investigated as an extension of the generalized linear model (McCullagh and Nelder, 1989)[22] and the *generalized varying coefficient model* (GVCM) (Hastie and Tibshirani 1993[4]; Cai, Fan, and Li 2000[2]).

Let Y be a response variable and its conditional expectation given by the covariate T and $\mathbf{X} = (X_1, X_2, \ldots, X_p)^\top$ is defined via a known link function g,

$$g\left(E(Y|\mathbf{X}, T)\right) = g(\mu) = \sum_{j \in S_1} \beta_j X_j + \sum_{j \in S_2} \gamma_j(T) X_j, \tag{1.1}$$

where S_1 and S_2 are mutually exclusive and complementary subsets of $\{1, \ldots, p\}$, $\{\beta_j, j \in S_1\}$ are the regression coefficients of covariates, and $\{\gamma_j(T), j \in S_2\}$ are unspecified smooth functions. The index variable T is a variable related to time or age in various research fields and it's interaction with other predictors is very important. We assume $T \in [0, 1]$ for simplicity. The variance of Y is a function of the mean, that is,

$$\mathrm{Var}(Y|\mathbf{X}, T) = V(\mu),$$

where $V(\cdot)$ is a known function. A lot of research work on variable selection and estimation for the GVCPLM has already been done such as Lam and Fan (2008)[12], Li and Liang (2008)[14], Lu (2008)[21], and Hu and Cui (2010)[5]. When the link function g is identity then the GCVPLM can be simplified into the semiparametric partially linear varying coefficient model (Li et al. 2002[13]; Xia, Zhang and Tong 2004[29]; Ahmad, Leelahanon and Li 2005[1]; Kai, Li and Zou 2011[11]; Li, Lin and Zhu 2012[16]).

It is worth mentioning that all the existing results are based on an important precondition that the linear part and the varying coefficient part are known in advance. However, this condition is usually unreasonable because in real application it is not clear whether the regression coefficients are dependent on the index variable or not. Statistically, treating constant coefficients as varying reduces estimation efficiency. Cai, Fan, and Li (2000)[2] determined constant coefficients based on the hypothesis testing. For the varying coefficient model, Xia, Zhang and Tong (2004)[29] developed a cross-validation procedure for judging constant and varying coefficients. Hu and Xia (2012)[6] developed a shrinkage method based on the adaptive Lasso to identify the constant coefficients. Structure's identification in both additive and varying-coefficient models have been studied by many researchers (Zhang, Cheng and Liu 2011[30]; Tang, et al. 2012[25]; Wang and Kulasekera 2012[26]; Lian, Chen and Yang 2012[18]; Lian, Liang and Ruppert 2015[20]; Lian, et al. 2014[19]). This work is motivated from the studies of Huang et al. (2012)[8] and Wang and Song (2013)[28]. Huang et al. (2012)[8] proposed a semiparametric model pursuit method for determining the linear and nonlinear effects in covariates for the partially linear model. Wang and Song (2013)[28] studied the identification of varying coefficients and constant coefficients for partially linear varying coefficient models using the group SCAD. Both of them used the profile least squares method to study the theory. Our work is a natural extension of Wang and Song (2013)[28] to more general types of responses using the quasi-likelihood. In addition, the proof of theocratical studies with quasi-likelihood is a challenging job.

Therefore, in this paper, we focused on the identification of the varying coefficients and constant coefficients of the GVCPLM, which can be embed into a GVCM. Using the spline method to estimate the varying coefficients, we transform the identification of the structure of the GVCPLM into a group variable selection problem. We apply

the group *smoothly clipped absolute deviation* (SCAD) penalty to identify the varying coefficients and the constant coefficients. Under some regular conditions, we prove that this method can consistently determine the structure of the GVCPLM, which means that it can accurately choose the varying coefficient components and the constant components. Furthermore, we demonstrate the convergence rate of the oracle estimator when the true model is known in advance. Simulation studies are conducted to evaluate the finite sample performance of the proposed method. The applicability of the proposed method is illustrated through a real data analysis on Burn Data (Cai, Fan and Li, 2000[2]).

The rest of the paper is organized as follows. Section 2 proposes the group SCAD approach and gives the theoretical results. Section 3 illustrates the performance of the proposed approach by simulation studies and a real data analysis. Some concluding remarks are presented in Section 4. The technical details are provided in the Appendix.

2. Identification for the GVCPLM via Penalty

The observed data for the ith subject or unit is $(\mathbf{X}_i, T_i, Y_i)$, $i = 1, 2, \ldots, n$, where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^\top$. The GVCPLM can be embedded into the GVCM (Hastie and Tibshirani, 1993[4])

$$g\left(E(Y_i|\mathbf{X}_i, T_i)\right) = g(\mu_i) = \sum_{j=1}^{p} \phi_j(T_i) X_{ij}. \tag{2.1}$$

If some of ϕ_j's are constants, then the model (2.1) becomes the GVCPLM (1.1). Therefore, our target is to determine which ϕ_j's are constants and which are not. We decompose ϕ_j into a constant component and a function component, that is, $\phi_j(T_i) = \beta_j + \gamma_j(T_i)$. For the unique identification of the β_j and $\gamma_j(T_i)$, we assume that $E\gamma_j(T_i) = 0$. Obviously, this decomposition is unique with $\beta_j = E\phi_j(T_i)$ and $\gamma_j(T_i) = \phi_j(T_i) - E\phi_j(T_i)$.

2.1. Spline Approximation. Let $0 = \xi_0 < \xi_1 < \cdots < \xi_{M_n+1} = 1$ be a partition of $[0, 1]$ into M_n+1 subintervals $I_k = [\xi_{k-1}, \xi_k), k = 1, \ldots, M_n$, and $I_{M_n+1} = [\xi_{M_n}, 1]$, where $M_n = O(n^\iota)$. Here, $0 < \iota < 0.5$ is a positive number such that $\max_{1\le k\le M_n+1} |\xi_k - \xi_{k-1}| = O(n^{-\iota})$. Denote $\mathcal{S}_m$ as the space of splines of degree m at all knots. The function s in this space possesses the following properties (Schumaker (1981)[24], page 108, definition 4.1):

(1) s is a polynomial of degree $m-1$ in any subinterval;
(2) If $m \ge 2$, s is $m-2$ times continuously differential on $[0, 1]$.

There exists a normalized B-spline basis function $\{B_k(t),\ 1 \le k \le q_n\}$ of $\mathcal{S}_m$, where $q_n = M_n + m$, is the dimension of the space. Hence, for any $\gamma_{nj} \in \mathcal{S}_m$, we have

$$\gamma_{nj}(t) = \sum_{k=1}^{q_n} \alpha_{jk} B_k(t), \quad 1 \le j \le p_n.$$

Under reasonable smoothness conditions, the function part γ_j can be approximated by the spline functions in $\mathcal{S}_m$. If $\alpha_{jk} = 0$ for $1 \le k \le q_n$, then ϕ_j is a constant. Thus the problem becomes variable selection of groups $\{\alpha_{jk}, 1 \le k \le q_n\}$.

2.2. Approach. Here, only the conditional mean and variance of the response are specified, so the usual likelihood is not available. We apply the (negative) quasi-likelihood function, which is defined by

$$Q(\mu, y) = \int_{\mu}^{y} \frac{y - s}{V(s)} \mathrm{d}s$$

and the negative quasi-likelihood of the collected data $\{(\mathbf{X}_i, T_i, Y_i),\ i = 1, 2, \ldots, n\}$ is

$$\ell_n(\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n) = \sum_{i=1}^{n} Q\left(g^{-1}\left(\sum_{j=1}^{p_n}\beta_{nj}X_{ij} + \sum_{j=1}^{p_n}\sum_{k=1}^{q_n}\alpha_{jk}B_k(T_i)X_{ij}\right), Y_i\right),$$

where $\boldsymbol{\beta}_n = (\beta_{n1}, \ldots, \beta_{np_n})^\top$, $\boldsymbol{\alpha}_n = (\boldsymbol{\alpha}_{n1}^\top, \ldots, \boldsymbol{\alpha}_{np_n}^\top)^\top$ with $\boldsymbol{\alpha}_{nj} = (\alpha_{j1}, \ldots, \alpha_{jq_n})^\top$. Here p_n indicates the dependence of p on n. Consider the penalized quasi-likelihood objective function

$$\ell_n(\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n) + n\sum_{j=1}^{p_n} p_\lambda(\|\boldsymbol{\alpha}_{nj}\|_{A_j}), \tag{2.2}$$

where $\|\boldsymbol{\alpha}_{nj}\|_{A_j} = (\boldsymbol{\alpha}_{nj}^\top A_j \boldsymbol{\alpha}_{nj})^{1/2}$ and $A_j = (a_{kl})_{q_n \times q_n}$ is a matrix with entries $a_{kl} = \int_0^1 B_k(t)B_l(t)\mathrm{d}t$. In this paper, we apply the SCAD penalty advocated by Fan and Li (2001), which is defined by the derivative $p'_\lambda(\theta) = \lambda\{I(\theta \le \lambda) + I(\theta > \lambda)(a\lambda - \theta)_+/[(a-1)\lambda]\}$ with $a > 2$ and $(t)_+ = tI(t > 0)$. As suggested by Fan and Li (2001)[3], we use $a = 3.7$. Denote λ by λ_n to emphasize the dependency of λ on n.

Define $\|\mathbf{s}\| = (\sum_{j=1}^d s_j^2)^{1/2}$ for any vector $\mathbf{s} \in \mathbf{R}^d$. For any function $f(t)$ on $[0,1]$, denote $\|f\|_2 = (\int_0^1 f^2(t)\mathrm{d}t)^{1/2}$, whenever the integral exists. For any square matrix $\mathbf{G}$, denote the smallest and largest eigenvalue of $\mathbf{G}$ as $\rho_{\min}(\mathbf{G})$ and $\rho_{\max}(\mathbf{G})$, respectively. For convenience, let $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$, where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip_n})^\top$. Denote $\mathbb{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^\top$, where $\mathbf{Z}_i = (\boldsymbol{Z}_{i1}^\top, \ldots, \boldsymbol{Z}_{ip_n}^\top)^\top$ is a $1 \times p_n q_n$ matrix and $\boldsymbol{Z}_{ij} = (B_1(T_i)X_{ij}, \ldots, B_{q_n}(T_i)X_{ij})^\top$. Then the objective function (2.2) can be rewritten as

$$\begin{aligned} & L(\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n; \lambda_n) \\ = & \sum_{i=1}^{n} Q\left(g^{-1}\left(\mathbf{X}_i^\top\boldsymbol{\beta}_n + \mathbf{Z}_i^\top\boldsymbol{\alpha}_n\right), Y_i\right) + n\sum_{j=1}^{p_n} p_{\lambda_n}(\|\boldsymbol{\alpha}_{nj}\|_{A_j}). \end{aligned} \tag{2.3}$$

The minimizer of (2.3) is defined by $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n) = \arg\min_{\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n} L(\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n; \lambda_n)$. Thus the estimator of $\gamma_j(t)$ is $\widehat{\gamma}_{nj}(t) = \sum_{k=1}^{q_n}\widehat{\alpha}_{jk}B_k(t)$, $1 \le j \le p_n$.

The set of indices of coefficients that are estimated to be constants in the regression model (1.1) is $\widehat{S}_1 \triangleq \{j\colon \|\widehat{\boldsymbol{\alpha}}_{nj}\|_{A_j} = 0\}$. Then we have

$$\widehat{\gamma}_{nj}(t) = 0,\ j \in \widehat{S}_1 \quad \text{and} \quad \widehat{\gamma}_{nj}(t) = \sum_{k=1}^{q_n}\widehat{\alpha}_{jk}B_k(t),\ j \notin \widehat{S}_1.$$

Hence, for $j \notin \widehat{S}_1$, $\widehat{\phi}_{nj}(t) = \widehat{\beta}_{nj} + \widehat{\gamma}_{nj}(t)$ and for $j \in \widehat{S}_1$, $\widehat{\phi}_{nj}(t) = \widehat{\beta}_{nj}$.

2.3. Asymptotic Results. Let μ be a non-negative integer and $\upsilon \in (0, 1]$ such that $r = \mu + \upsilon > 0.5$. Let $\mathcal{H}$ be the class of functions h on $[0,1]$ whose μ-th derivative $h^{(\mu)}$ exists and satisfies a Lipschitz condition of order υ, $|h^{(\mu)}(t_1) - h^{(\mu)}(t_2)| \le C|t_1 - t_2|^\upsilon$, for $0 \le t_1, t_2 \le 1$, where C is a positive constant. The order of the polynomial spline $m \ge \mu + 1$.

In order to study the consistency of the identification for the GVCPLM and the convergence rate of estimators of varying coefficients, we first define the oracle estimator. Denote the true value of the varying coefficient $\phi_j(t)$ by $\phi_{0j}(t)$, and $\phi_{0j}(t) = \beta_{0j} + \gamma_{0j}(t)$. Let $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p_n})^\top$ and $\boldsymbol{\gamma}_0(T_i) = (\gamma_{01}(T_i), \ldots, \gamma_{0p_n}(T_i))^\top$, for $i = 1, 2, \ldots, n$. We have $\|\gamma_{0j}\|_2 = 0$ for $j \in S_1$. Let k_n be the cardinality of S_1, the number of linear components in the model. Let d_n be the cardinality of S_2, the number of varying coefficients in the model with $k_n + d_n = p_n$. Suppose that $\gamma_{0j}(t) \in \mathcal{H}, j \in S_2$. For $j \in S_2$, according

to the condition (A1) and Corollary 6.21 in Schumaker (1981)[24], there exists a vector $\boldsymbol{\alpha}_{n0j} = (\alpha_{0j1}, \ldots, \alpha_{0jq_n})^\top$ satisfying that $\|\gamma_{0j}(t) - \sum_{k=1}^{q_n} \alpha_{0jk} B_k(t)\|_2 = O(q_n^{-r})$. Define

$$(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\alpha}}_n) = \underset{\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n}{\arg\min} \left\{ \sum_{i=1}^{n} Q\left(g^{-1}\left(\mathbf{X}_i^\top \boldsymbol{\beta}_n + \mathbf{Z}_i^\top \boldsymbol{\alpha}_n\right), Y_i\right) : \boldsymbol{\alpha}_{nj} = \mathbf{0}, j \in S_1 \right\}.$$

They are the oracle estimators of $\boldsymbol{\beta}_n$ and $\boldsymbol{\alpha}_n$ assuming the constant coefficients were known in advance. Let $\mathbf{Z}_{2i} = (\mathbf{Z}_{ij}, j \in S_2)^\top$, $\boldsymbol{\alpha}_{2n} = (\boldsymbol{\alpha}_{nj}^\top, j \in S_2)^\top$ and $\boldsymbol{\alpha}_{20} = (\boldsymbol{\alpha}_{n0j}^\top, j \in S_2)^\top$. Denote $\mathbb{Z}_2 = (\mathbf{Z}_{2i}, i = 1, \ldots, n)^\top$. We should note that the oracle estimator can not be obtained in application. Here, we use it to study the theoretical properties. Hence, the oracle estimator of $\gamma_j(t)$ is

$$\widetilde{\gamma}_{nj}(t) = 0, j \in S_1 \quad \text{and} \quad \widetilde{\gamma}_{nj}(t) = \sum_{k=1}^{q_n} \widetilde{\alpha}_{jk} B_k(t), j \in S_2.$$

Write $\widetilde{\phi}_{nj}(t) = \widetilde{\beta}_{nj} + \widetilde{\gamma}_{nj}(t)$ for $j \in S_2$.

For simplicity, define $q_l(x, y) = \partial^l / \partial x^l Q(g^{-1}(x), y)$ for $l = 1, 2, 3$. So we have

$$q_1(x, y) = -(y - g^{-1}(x))\rho_1(x) \quad \text{and} \quad q_2(x, y) = \rho_2(x) - (y - g^{-1}(x))\rho_1'(x),$$

where $\rho_l(x) = [dg^{-1}(x)/dx]^l / V(g^{-1}(x))$. To give the asymptotic results, the following regularity assumptions are used.

- (A1) The covariate T has a continuous density and there exist constants C_1 and C_2 such that the density function f of T satisfies $0 < C_1 \le f(t) \le C_2 < \infty$ for all $t \in [0, 1]$.
- (A2) The covariates $|X_{ij}| \, (1 \le i \le n, 1 \le j \le p_n)$ are bounded away from zero and infinity.
- (A3) The eigenvalues of $\mathbb{X}^\top \mathbb{X}/n$ are bounded away from zero and infinity. The eigenvalues of $\sum_{i=1}^n q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{Z}_i^\top \boldsymbol{\alpha}_0, Y_i) \mathbf{X}_i \mathbf{X}_i^\top / n$ are bounded away from zero and infinity.
- (A4) $E[q_1(\mathbf{X}_1^\top \boldsymbol{\beta}_0 + \mathbf{Z}_1^\top \boldsymbol{\alpha}_0, Y_1)]^2 < \infty$, $E[q_2(\mathbf{X}_1^\top \boldsymbol{\beta}_0 + \mathbf{Z}_1^\top \boldsymbol{\alpha}_0, Y_1)]^2 < \infty$ and $c < q_2(\mathbf{X}_1^\top \boldsymbol{\beta}_0 + \mathbf{X}_1^\top \boldsymbol{\gamma}_0(T_1), Y_1) < C$ for some constants $C > c > 0$.
- (A5) $E[q_3(\cdot, Y_1)]^2$ is bounded.
- (A6) $p_n^2/n \to 0$, $d_n^2 q_n^2/n \to 0$, $p_n d_n^2 q_n^{-2r} \to 0$ and $d_n^3 q_n^{-2r+1} \to 0$.
- (A7) $\sqrt{(p_n + d_n q_n)/n} + d_n q_n^{-r} \ll \lambda_n \ll \inf_{j \in S_2} \|\gamma_{0j}(t)\|_2$, where two sequences $a_n \ll b_n$ means $a_n = o(b_n)$.

Conditions (A1-A5) are regular conditions in the studies about semiparametric statistical inference (Lam and Fan, 2008[12]; Huang, Horowitz and Wei 2010[7]; Lian, 2012[17]). From the condition (A3), it is easy to know that all eigenvalues of $q_n \mathbb{Z}^\top \mathbb{Z}/n$ are bounded away from zero and infinity (Lemma A.1 in Huang, Wu and Zhou, 2004[9]). Conditions (A6) and (A7) seem to be complex, we discuss them carefully. Suppose that an optimal $q_n = O(n^{1/(2r+1)})$, $p_n = O(n^{c_1})$ and $d_n = O(n^{c_2})$ for some $0 < c_2 < c_1 < 1$. Then the condition (A6) becomes $2c_1 < 1$, $c_1 + 2c_2 < 2r/(2r+1)$ and $3c_2 < (2r-1)/(2r+1)$. If $\inf_{j \in S_2} \|\gamma_{0j}(t)\|_2 \ge M n^{-c_3}$ for some constants $M > 0$ and $c_3 \ge 0$, $\lambda_n = n^{-c_4}$ with $c_4 > 0$, then the condition (A7) becomes

$$n^{(c_1-1)/2} + n^{c_2 - r/(2r+1)} \ll n^{-c_4} \ll n^{-c_3},$$

which is equivalent to $c_3 < c_4 < \min\{(1 - c_1)/2, r/(2r+1) - c_2\}$.

2.1. Theorem. *Suppose that conditions (A1-A6) hold, then we have*

- (i) $\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\| = O_P(\sqrt{(p_n + d_n q_n)/n} + d_n q_n^{-r})$.
- (ii) $\|\widetilde{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}_0\| = O_P(\sqrt{(p_n + d_n q_n)/n} + d_n q_n^{-r})$.

Furthermore,

$$\sum_{j\in S_2}\|\widetilde{\phi}_{nj}(t)-\phi_{0j}(t)\|_2^2 \quad = \quad O_P\left((p_n+d_nq_n)/n+d_n^2q_n^{-2r}\right).$$

Proof. Let $\delta_n=\sqrt{p_n/n}+p_nq_n^{-r}$. It suffices to show that for any given $\epsilon>0$, there exists a large constant C such that

$$\Pr\left\{\inf_{\|\mathbf{u}\|=C}\ell_n(\boldsymbol{\beta}_0+\delta_n\mathbf{u}_1,\boldsymbol{\alpha}_{20}+\delta_n\mathbf{u}_2)>\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_{20})\right\}\geq 1-\epsilon, \tag{2.4}$$

where $\mathbf{u}=(\mathbf{u}_1^\top,\mathbf{u}_2^\top)^\top$ with $\mathbf{u}_1$ being $p_n\times 1$ vector and $\mathbf{u}_2$ being $d_n\times 1$ vector . (2.4) implies that with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{\boldsymbol{\beta}_0+\delta_n\mathbf{u}_1:\|\mathbf{u}_1\|\leq C\}$. That is, there exists a local minimizer such that $\|\widetilde{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\|=O_P(\sqrt{p_n/n}+p_nq_n^{-r})$. Due to the convexity of $\ell_n(\cdot)$, $\widetilde{\boldsymbol{\beta}}_n$ is the global minimizer. Similarly, $\|\widetilde{\boldsymbol{\alpha}}_{2n}-\boldsymbol{\alpha}_{20}\|=O_P(\sqrt{p_n/n}+p_nq_n^{-r})$.

Let $\boldsymbol{\theta}_0=(\boldsymbol{\beta}_0^\top,\boldsymbol{\alpha}_{20}^\top)^\top$, $\mathbf{W}_i=(\mathbf{X}_i^\top,\mathbf{Z}_{2i}^\top)^\top$ and $m_{0i}=\mathbf{X}_i^\top\boldsymbol{\beta}_0+\mathbf{Z}_{2i}^\top\boldsymbol{\alpha}_{20}=\mathbf{W}_i^\top\boldsymbol{\theta}_0$, then $\mathbb{W}=(\mathbb{X},\mathbb{Z}_2)$ and $\boldsymbol{m}_0=\mathbb{W}^\top\boldsymbol{\theta}_0$. By some calculation, we have

$$\begin{aligned}
&\ell_n(\boldsymbol{\beta}_0+\delta_n\mathbf{u}_1,\boldsymbol{\alpha}_{20}+\delta_n\mathbf{u}_2)-\ell_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}_{20})\\
&=\sum_{i=1}^n Q\left(g^{-1}\left(\mathbf{X}_i^\top(\boldsymbol{\beta}_0+\delta_n\mathbf{u}_1)+\mathbf{Z}_{2i}^\top(\boldsymbol{\alpha}_{20}+\delta_n\mathbf{u}_2)\right),Y_i\right)\\
&\quad-\sum_{i=1}^n Q\left(g^{-1}\left(\mathbf{X}_i^\top\boldsymbol{\beta}_0+\mathbf{Z}_{2i}^\top\boldsymbol{\alpha}_{20}\right),Y_i\right)\\
&=\delta_n\sum_{i=1}^n q_1(m_{0i},Y_i)(\mathbf{X}_i^\top\mathbf{u}_1+\mathbf{Z}_{2i}^\top\mathbf{u}_2)+\frac{\delta_n^2}{2}\sum_{i=1}^n q_2(m_{0i},Y_i)(\mathbf{X}_i^\top\mathbf{u}_1+\mathbf{Z}_{2i}^\top\mathbf{u}_2)^2\\
&\quad+\frac{\delta_n^3}{6}\sum_{i=1}^n q_3(m_i^*,Y_i)(\mathbf{X}_i^\top\mathbf{u}_1+\mathbf{Z}_{2i}^\top\mathbf{u}_2)^3\\
&\triangleq \mathrm{I}_{n1}+\mathrm{I}_{n2}+\mathrm{I}_{n3},
\end{aligned}$$

where m_i^* is between m_{0i} and $m_{0i}+\delta_n\mathbf{W}_i^\top\mathbf{u}$.

For I_{n1}, let $\boldsymbol{q}_1(\boldsymbol{m}_0,\boldsymbol{Y})=(q_1(m_{01},Y_1),\ldots,q_1(m_{0n},Y_n))^\top$, then

$$\begin{aligned}
\|\mathrm{I}_{n1}\|^2 &= \delta_n^2\|\boldsymbol{q}_1(\boldsymbol{m}_0,\boldsymbol{Y})^\top\mathbb{W}\mathbf{u}\|^2\\
&\leq \delta_n^2\|P_W\boldsymbol{q}_1(\boldsymbol{m}_0,\boldsymbol{Y})\|^2\|\mathbb{W}\mathbf{u}\|^2,
\end{aligned}$$

where $P_W=\mathbb{W}(\mathbb{W}^\top\mathbb{W})^{-1}\mathbb{W}^\top$ is a projection matrix. By the condition (A3), it is easy to show that all the eigenvalues of $\mathbb{W}^\top\mathbb{W}/n$ are bounded away from zero and infinity. Hence, $\|\mathbb{W}\mathbf{u}\|^2=O(n\|\mathbf{u}\|^2)$. In addition,

$$\begin{aligned}
&\|P_W\boldsymbol{q}_1(\boldsymbol{m}_0,\boldsymbol{Y})\|^2\\
&=\|P_W[\boldsymbol{q}_1(\mathbb{X}\boldsymbol{\beta}_0+\mathbb{X}\boldsymbol{\gamma}_0(\boldsymbol{T}),\boldsymbol{Y})+\boldsymbol{q}_1(\mathbb{X}\boldsymbol{\beta}_0+\mathbb{Z}\boldsymbol{\alpha}_0,\boldsymbol{Y})-\boldsymbol{q}_1(\mathbb{X}\boldsymbol{\beta}_0+\mathbb{X}\boldsymbol{\gamma}_0(\boldsymbol{T}),\boldsymbol{Y})]\|^2\\
&\leq 2\|P_W\boldsymbol{q}_1(\mathbb{X}\boldsymbol{\beta}_0+\mathbb{X}\boldsymbol{\gamma}_0(\boldsymbol{T}),\boldsymbol{Y})\|^2\\
&\quad+2\|P_W[\boldsymbol{q}_1(\mathbb{X}\boldsymbol{\beta}_0+\mathbb{Z}\boldsymbol{\alpha}_0,\boldsymbol{Y})-\boldsymbol{q}_1(\mathbb{X}\boldsymbol{\beta}_0+\mathbb{X}\boldsymbol{\gamma}_0(\boldsymbol{T}),\boldsymbol{Y})]\|^2\\
&\triangleq \mathrm{J}_{n1}+\mathrm{J}_{n2},
\end{aligned}$$

where $\mathbb{X}\boldsymbol{\gamma}_0(\boldsymbol{T})=\left(\mathbf{X}_1^\top\boldsymbol{\gamma}_0(T_1),\ldots,\mathbf{X}_n^\top\boldsymbol{\gamma}_0(T_n)\right)^\top$. By the condition (A4), we have

$$\mathrm{J}_{n1}=O_P(tr(P_W))=O_P(p_n+d_nq_n).$$

For J_{n2}, let $\mathbf{B}(T_i) = Diag(\boldsymbol{B}_1(T_i)^\top, \ldots, \boldsymbol{B}_{p_n}(T_i)^\top)$ with $\boldsymbol{B}_j(T_i) = (B_{j1}(T_i), \ldots, B_{jq_n}(T_i))^\top$

$$\begin{aligned}
\mathrm{J}_{n2} &\leq 2\sum_{i=1}^{n}\left[q_1(\mathbf{X}_i^\top\boldsymbol{\beta}_0 + \mathbf{Z}_i^\top\boldsymbol{\alpha}_0, Y_i) - q_1(\mathbf{X}_i^\top\boldsymbol{\beta}_0 + \mathbf{X}_i^\top\boldsymbol{\gamma}_0(T_i), Y_i)\right]^2 \\
&\leq 4\sum_{i=1}^{n} q_2^2(\mathbf{X}_i^\top\boldsymbol{\beta}_0 + \mathbf{X}_i^\top\boldsymbol{\gamma}_0(T_i), Y_i)\left[\mathbf{X}_i^\top(\boldsymbol{\gamma}_0(T_i) - \mathbf{B}(T_i)\boldsymbol{\alpha}_0)\right]^2 \\
&\quad +4\sum_{i=1}^{n} q_3^2(\tilde{m}_i, Y_i)\left[\mathbf{X}_i^\top(\boldsymbol{\gamma}_0(T_i) - \mathbf{B}(T_i)\boldsymbol{\alpha}_0)\right]^4 \\
&\quad = O_P(nd_n^2 q_n^{-2r}),
\end{aligned}$$

where $\tilde{m}_i$ is between $\mathbf{X}_i^\top\boldsymbol{\beta}_0 + \mathbf{Z}_i^\top\boldsymbol{\alpha}_0$ and $\mathbf{X}_i^\top\boldsymbol{\beta}_0 + \mathbf{X}_i^\top\boldsymbol{\gamma}_0(T_i)$. In fact, by the condition (A4)

$$\begin{aligned}
&E\left\{\sum_{i=1}^{n} q_2^2(\mathbf{X}_i^\top\boldsymbol{\beta}_0 + \mathbf{X}_i^\top\boldsymbol{\gamma}_0(T_i), Y_i)\left[\mathbf{X}_i^\top(\boldsymbol{\gamma}_0(T_i) - \mathbf{B}(T_i)\boldsymbol{\alpha}_0)\right]^2\right\} \\
&\leq M\sum_{i=1}^{n}\|\mathbf{X}_i\|^2 E\|\boldsymbol{\gamma}_0(T_i) - \mathbf{B}(T_i)\boldsymbol{\alpha}_0\|^2 \\
&= O(nd_n^2 q_n^{-2r}),
\end{aligned}$$

and

$$\begin{aligned}
&E\left\{\sum_{i=1}^{n} q_3^2(\tilde{m}_i, Y_i)\left[\mathbf{X}_i^\top(\boldsymbol{\gamma}_0(T_i) - \mathbf{B}(T_i)\boldsymbol{\alpha}_0)\right]^4\right\} \\
&\leq M\sum_{i=1}^{n}\|\mathbf{X}_i\|^4 E\|\boldsymbol{\gamma}_0(T_i) - \mathbf{B}(T_i)\boldsymbol{\alpha}_0\|^4 \\
&= O(nd_n^4 q_n^{-4r}) = o(nd_n^2 q_n^{-2r})
\end{aligned}$$

Hence, $|\mathrm{I}_{n1}| = O_P\left(n\delta_n(\sqrt{(p_n + d_n q_n)/n} + d_n q_n^{-r})\right)\|\mathbf{u}\|$.

For I_{n2}, by the condition (A3), we have

$$\begin{aligned}
\mathrm{I}_{n2} &\geq 2\delta_n^2\sum_{i=1}^{n} q_2(m_{0i}, Y_i)(\mathbf{W}_i^\top\mathbf{u})^2 \\
&\geq Mn\delta_n^2\|\mathbf{u}\|^2.
\end{aligned}$$

For I_{n3}, by the condition (A6), we have

$$\begin{aligned}
|\mathrm{I}_{n3}| &\leq \delta_n^3\sum_{i=1}^{n}|q_3(m_i^*, Y_i)|\cdot|\mathbf{W}_i^\top\mathbf{u}|^3 \\
&\leq M\delta_n^3(\sqrt{p_n + d_n q_n})\|\mathbf{u}\|\sum_{i=1}^{n}(\mathbf{W}_i^\top\mathbf{u})^2 \\
&\leq Mn\delta_n^3(\sqrt{p_n + d_n q_n})\|\mathbf{u}\|^3.
\end{aligned}$$

Hence, I_{n2} dominates all of the items uniformly when a sufficiently large C is chosen. As I_{n2} is positive, this completes the proof of Theorem 2.1.

Next, we prove

$$\begin{aligned}
& \sum_{j\in S_2} \|\widetilde{\gamma}_{nj}(t) - \gamma_{0j}(t)\|_2^2 \\
\leq\ & 2\sum_{j\in S_2} \|\widetilde{\gamma}_{nj}(t) - \boldsymbol{B}_j(t)^\top \boldsymbol{\alpha}_{0j}\|_2^2 + 2\sum_{j\in S_2} \|\boldsymbol{B}_j(t)^\top \boldsymbol{\alpha}_{0j} - \gamma_{0j}(t)\|_2^2 \\
=\ & O_P\left((q_n^{-1}(p_n + d_n q_n)/n + d_n^2 q_n^{-2r-1}) + d_n q_n^{-2r}\right) \\
=\ & O_P\left((p_n + d_n q_n)/n + d_n^2 q_n^{-2r}\right).
\end{aligned}$$

Hence, $\sum_{j\in S_2} \|\widetilde{\phi}_{nj}(t) - \phi_{0j}(t)\|_2^2 = O_P\left((p_n + d_n q_n)/n + d_n^2 q_n^{-2r}\right)$. □

This theorem gives the convergence rate of the oracle estimator. If p_n is bounded, each component in (2.1) is second order differentiable ($r = 2$) and take $q_n = O(n^{1/5})$, then the convergence rate in Theorem 2.1 is $n^{-4/5}$, which is the optimal convergence rate in nonparametric regression.

2.2. Lemma. *Suppose that $1/(q_n(a-1))$ is less than the smallest eigenvalue of $\sum_{i=1}^n q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{Z}_i^\top \boldsymbol{\alpha}_0, Y_i)\mathbf{Z}_i\mathbf{Z}_i^\top/n$, so we have $(\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n)$ is the solution of (2.3) if and only if*

(1) $\sum_{i=1}^n q_1(\mathbf{X}_i^\top \boldsymbol{\beta}_n + \mathbf{Z}_i^\top \boldsymbol{\alpha}_n, Y_i)X_{ij} = 0, \quad$ for $j = 1, 2, \ldots, p_n$,
(2) $\sum_{i=1}^n q_1(\mathbf{X}_i^\top \boldsymbol{\beta}_n + \mathbf{Z}_i^\top \boldsymbol{\alpha}_n, Y_i)\boldsymbol{Z}_{ij} = 0$ and $\|\boldsymbol{\alpha}_{nj}\|_{A_j} \geq a\lambda_n, \quad$ for $j \in S_2$,
(3) $\|\sum_{i=1}^n q_1(\mathbf{X}_i^\top \boldsymbol{\beta}_n + \mathbf{Z}_i^\top \boldsymbol{\alpha}_n, Y_i)\boldsymbol{Z}_{ij}\| \leq n\lambda_n$ and $\|\boldsymbol{\alpha}_{nj}\|_{A_j} < \lambda_n, \quad$ for $j \in S_1$.

This lemma is a direct extension of Theorem 1 in Kim *et al.* (2008)[10] to the case of quasi-likelihood. Thus, we omit the proof of this lemma.

2.3. Theorem. *(Selection Consistency). Suppose that $1/(q_n(a-1))$ is less than the smallest eigenvalue of $\sum_{i=1}^n q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{Z}_i^\top \boldsymbol{\alpha}_0, Y_i)\mathbf{Z}_i\mathbf{Z}_i^\top/n$. Under conditions (A1)-(A7), we have*

$$\Pr(\widehat{\boldsymbol{\beta}}_n = \widetilde{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\alpha}}_n = \widetilde{\boldsymbol{\alpha}}_n) \to 1.$$

Consequently, $\Pr(\widehat{S}_1 = S_1) \to 1$.

Proof. Since $1/(q_n(a-1))$ is less than the smallest eigenvalue of $\sum_{i=1}^n q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{Z}_i^\top \boldsymbol{\alpha}_0, Y_i)\mathbf{Z}_i\mathbf{Z}_i^\top/n$, the objective function (2.3) is a convex function, so we only need to show that $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\alpha}}_n)$ satisfies equations (1)-(3) of Lemma 2.2. By the definition of $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\alpha}}_n)$, it is easy to know (1) holds, and $\sum_{i=1}^n q_1(\mathbf{X}_i^\top \widetilde{\boldsymbol{\beta}}_n + \mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n, Y_i)\boldsymbol{Z}_{ij} = 0 \quad$ for $j \in S_2$. Next, we verify $\|\widetilde{\boldsymbol{\alpha}}_{nj}\|_{A_j} \geq a\lambda_n, \quad$ for $j \in S_2$. In fact,

$$\begin{aligned}
\|\widetilde{\boldsymbol{\alpha}}_{nj}\|_{A_j} &= \|\widetilde{\gamma}_{nj}(t) - \gamma_{0j}(t) + \gamma_{0j}(t)\|_2 \\
&\geq \min_{j\in S_2} \|\gamma_{0j}(t)\|_2 - \|\widetilde{\gamma}_{nj}(t) - \gamma_{0j}(t)\|_2.
\end{aligned}$$

By the condition (A7), we have $\min_{j\in S_2} \|\gamma_{0j}(t)\|_2 \gg \lambda_n$ and $\|\widetilde{\gamma}_{nj}(t) - \gamma_{0j}(t)\|_2 \ll \lambda_n$, so (2) holds.

Since $\|\widetilde{\boldsymbol{\alpha}}_{nj}\|_{A_j} = 0$, for $j \in S_1$, so $\|\boldsymbol{\alpha}_{nj}\|_{A_j} < \lambda_n, \quad$ for $j \in S_1$. Furthermore,

$$\begin{aligned}
&\sum_{i=1}^{n} q_1(\mathbf{X}_i^\top \widetilde{\boldsymbol{\beta}}_n + \mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n, Y_i)\boldsymbol{Z}_{ij} \\
&= \sum_{i=1}^{n} q_1(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)\boldsymbol{Z}_{ij} \\
&\quad + \sum_{i=1}^{n} q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)\boldsymbol{Z}_{ij}(\mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n - \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i)) \\
&\quad + \frac{1}{2}\sum_{i=1}^{n} q_3(\cdot, Y_i)\boldsymbol{Z}_{ij}(\mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n - \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i))^2 \\
&\triangleq \mathrm{II}_{n1} + \mathrm{II}_{n2} + \mathrm{II}_{n3}.
\end{aligned}$$

First, for II_{n1}, by conditions (A2) and (A4), and Lemma 6.1 of Zhou, Shen and Wolf (1998)[31], we have

$$\begin{aligned}
&\Pr\left(\max_{j\in S_1}\left\|\sum_{i=1}^{n} q_1(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)\boldsymbol{Z}_{ij}\right\| > n\lambda_n/3\right) \\
&\le \frac{9k_n}{n^2\lambda_n^2} E\left\|\sum_{i=1}^{n} q_1(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)\boldsymbol{Z}_{ij}\right\|^2 \\
&\le \frac{9k_n}{n^2\lambda_n^2}\sum_{k=1}^{q_n}\left\{\sum_{i=1}^{n} E\left[q_1^2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)X_{ij}^2 B_{jk}^2(T_i)\right]\right\} \\
&= \frac{9k_n}{n^2\lambda_n^2} O(n) = O\left(p_n/(n\lambda_n^2)\right) = o(1),
\end{aligned}$$

For II_{n2}, by conditions (A2), (A3) and (A4),

$$\begin{aligned}
&\left\|\sum_{i=1}^{n} q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)\boldsymbol{Z}_{ij}(\mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n - \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i))\right\|^2 \\
&= \sum_{k=1}^{q_n}\left[\sum_{i=1}^{n} q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)B_{jk}(T_i)X_{ij}\mathbf{X}_i^\top(\mathbf{B}(T_i)\widetilde{\boldsymbol{\alpha}}_n - \boldsymbol{\gamma}_0(T_i))\right]^2 \\
&\le \sum_{k=1}^{q_n}\left[\sum_{i=1}^{n} q_2^2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)B_{jk}^2(T_i)X_{ij}^2\right]\left[\sum_{i=1}^{n}(\mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n - \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i))^2\right] \\
&= O_P(n)O_P\left(n((p_n + d_n q_n)/n + d_n^2 q_n^{-2r})\right)
\end{aligned}$$

By the condition (A7),

$$\left\|\sum_{i=1}^{n} q_2(\mathbf{X}_i^\top \boldsymbol{\beta}_0 + \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i), Y_i)\boldsymbol{Z}_{ij}(\mathbf{Z}_i^\top \widetilde{\boldsymbol{\alpha}}_n - \mathbf{X}_i^\top \boldsymbol{\gamma}_0(T_i))\right\| = o_p(n\lambda_n)$$

Similarly, we obtain $\mathrm{II}_{n3} = o_P(n\lambda_n)$. Hence (3) holds. □

This theorem shows that the proposed estimator can correctly distinguish constant effects and varying effects with probability approaching 1. Hence, the proposed estimator enjoys the oracle property in the sense that it is the same as the oracle estimator assuming the identity of constants and varying coefficients were known in advance.

2.4. Computation. To solve the minimizer of (2.2), we use the local quadratic approximation (LQA) proposed by Fan and Li (2001)[3]. For a given $u_0 \neq 0$, the penalty function can be locally approximated by a quadratic function as

$$[p_{\lambda_n}(|u|)]' = p'_{\lambda_n}(|u|)\text{sgn}(u) \approx \{p'_{\lambda_n}(|u_0|)/|u_0|\}u,$$

namely,

$$p_{\lambda_n}(|u|) \approx p_{\lambda_n}(|u_0|) + \frac{1}{2}\{p'_{\lambda_n}(|u_0|)/|u_0|\}(u^2 - u_0^2).$$

More specifically, we take the initial value $\boldsymbol{\alpha}_{nj}^0$ with $\|\boldsymbol{\alpha}_{nj}^0\|_{A_j} > 0, j = 1, \ldots, p_n$. The penalty term can be approximated as

$$p_{\lambda_n}(\|\boldsymbol{\alpha}_{nj}\|_{A_j}) \approx p_{\lambda_n}(\|\boldsymbol{\alpha}_{nj}^0\|_{A_j}) + \frac{1}{2}\{p'_{\lambda_n}(\|\boldsymbol{\alpha}_{nj}^0\|_{A_j})/\|\boldsymbol{\alpha}_{nj}^0\|_{A_j}\}(\|\boldsymbol{\alpha}_{nj}\|_{A_j}^2 - \|\boldsymbol{\alpha}_{nj}^0\|_{A_j}^2).$$

Hence, removing the irrelevant constants, (2.2) becomes

$$L(\boldsymbol{\beta}_n, \boldsymbol{\alpha}_n; \lambda_n) = \sum_{i=1}^{n} Q\left(g^{-1}\left(\mathbf{X}_i^\top \boldsymbol{\beta}_n + \mathbf{Z}_i^\top \boldsymbol{\alpha}_n\right), Y_i\right) + n/2\boldsymbol{\alpha}_n^\top \Sigma_{\lambda_n}(\boldsymbol{\alpha}_n^0)\boldsymbol{\alpha}_n,$$

where

$$\Sigma_{\lambda_n}(\boldsymbol{\alpha}_n^0) = \text{diag}\left\{\frac{p'_{\lambda_n}(\|\boldsymbol{\alpha}_{nj}^0\|_{A_j})}{\|\boldsymbol{\alpha}_{nj}^0\|_{A_j}} A_j, j = 1, \ldots, p_n\right\}.$$

The Newton-Raphson iterative algorithm is used to find the solution.

Several tuning parameter's selection procedures are available in the literature such as cross-validation, generalized cross-validation, AIC and BIC. As Wang, Li and Tsai (2007)[27],Huang, Wei and Ma (2012)[8], and Li, Xue and Lian (2012)[15], we use BIC (Schwarz, 1978[23]) to select the tuning parameter for each method. The BIC is defined by

$$BIC(\lambda) = 2\ell_n(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\alpha}}_n) + q_n df_\lambda \log n,$$

where $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\alpha}}_n)$ is the minimizer of the equation (2.2) for a given λ and df_λ is the number of varying coefficients.

3. Numerical Examples

To examine the finite sample performance of the proposed method, we conduct simulation studies for the logistic regression model and Poisson regression model. We also present an analysis of a real data.

Example 1. In this example, we generate the data from the following model

$$\text{logit}[\Pr(Y_i = 1|\boldsymbol{x}_i, T_i)] \quad = \quad \sum_{j=1}^{p_n} \beta_j(T_i)X_{ij}, \quad i = 1, 2, \ldots, n,$$

where $\boldsymbol{x}_i = (X_{ij}, j = 1, \ldots, p_n)^\top$ with X_{ij} being generated from standard normal distribution $N(0, 1)$. The index variable T_i is uniformly distributed over $[0, 1]$. We consider $p_n = 4$ and 6. When $p_n = 4$, the true coefficients $\beta_1(T_i) = 2, \beta_2(T_i) = 1, \beta_3(T_i) = 2T_i^3 + 4T_i^2 - 4T_i$ and $\beta_4(T_i) = 2\sin(2\pi T_i)$. Thus, the number of constant coefficients is 2 and the number of varying coefficients is 2. When $p_n = 6$, the true coefficients $\beta_1(T_i) = 2, \beta_2(T_i) = 1, \beta_3(T_i) = -2, \beta_4(T_i) = -1, \beta_5(T_i) = 2T_i^3 + 4T_i^2 - 4T_i$ and $\beta_6(T_i) = 2\sin(2\pi T_i)$. That is, the number of constant coefficients is 4 and the number of varying coefficients is 2. In the following simulation, we use cubic B-spline to approximate each function. As suggested by Lian (2012)[17], we fix the number of basis functions with $q_n = 5$. The sample size n is set to be 500, 600 and 700. For each case, 100 replications are conducted.

The simulation results are presented in Tables 1–3. We compare the performance of the SCAD and Lasso. Table 1 shows the number of times each component selected as a varying coefficient based on 100 replications. Table 2 gives the average number of varying coefficients being selected (NVC) and the average number of true varying coefficients being selected (NTVC). In Table 3, we present the root mean squared errors for all component functions, which is defined by

$$RMSE(j) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}[\hat{\beta}_j(t_i) - \beta_j(t_i)]^2}, \quad j = 1, \ldots, p_n,$$

where $\{t_i, i = 1, \ldots, 1000\}$ is a grid equally spaced on $[0, 1]$. Enclosed in parentheses are the corresponding standard errors.

From Table 1 and Table 2, we can make the following observations: Table 1 shows that the group SCAD was more accurate than the group Lasso in identifying the varying coefficients and the constant coefficients. Table 2 indicates that the group SCAD seems to select less number of varying coefficients than the Lasso especially for the smaller sample size. The reason may be that the Lasso always tends to keep more variables. In Table 3, we can see that, the proposed method with the group SCAD has the smaller mean square errors than the group Lasso for the constant coefficients, and is similar with the oracle estimator. It is valuable to note that both the group SCAD estimators and the group Lasso estimators of the varying coefficients could outperform the oracle estimators for the varying coefficients. This can be explained by that some shrinkage is beneficial to reducing overfitting and improving stability. Overall, the proposed method with the group SCAD is effective in distinguishing the constant coefficients from the varying coefficients in the simulation models. Figure 1 shows the estimated varying coefficient functions along with the true function components from the group SCAD and group Lasso. The estimated coefficient functions are computed based on the mean of $\widehat{\boldsymbol{\alpha}}_n$ in 200 runs when $n = 700$ and $p_n = 4$. We can see that the SCAD performs better than the Lasso.

Table 1. Number of times each component was selected as a varying coefficient in the 100 replications for the logistic model.

p_n	n	Method	β_1	β_2	β_3	β_4	β_5	β_6
4	500	SCAD	5	2	48	100		
		Lasso	18	24	89	98		
	600	SCAD	4	0	54	100		
		Lasso	22	27	95	100		
	700	SCAD	2	2	58	100		
		Lasso	21	31	98	100		
6	500	SCAD	3	1	4	0	30	91
		Lasso	2	6	4	4	65	83
	600	SCAD	2	1	2	0	31	97
		Lasso	3	5	1	6	76	89
	700	SCAD	0	0	0	0	46	99
		Lasso	2	5	2	7	88	97

Table 2. Model selection results for the logistic model

p_n	n	Method	NVC	NTVC
4	500	SCAD	1.55 (0.657)	1.48 (0.502)
		Lasso	2.29 (0.913)	1.87 (0.393)
	600	SCAD	1.58 (0.572)	1.54 (0.501)
		Lasso	2.44 (0.891)	1.95 (0.219)
	700	SCAD	1.62 (0.565)	1.58 (0.496)
		Lasso	2.50 (0.847)	1.98 (0.141)
6	500	SCAD	1.29 (0.795)	1.21 (0.591)
		Lasso	1.64 (1.000)	1.48 (0.772)
	600	SCAD	1.33 (0.682)	1.28 (0.514)
		Lasso	1.80 (0.841)	1.65 (0.672)
	700	SCAD	1.45 (0.520)	1.45 (0.520)
		Lasso	2.01 (0.703)	1.85 (0.435)

Table 3. The root of average mean square error of each component for the logistic model.

p_n	n		SCAD	Lasso	Oracle
4	500	β_1	0.218 (0.272)	0.363 (0.223)	0.157 (0.126)
		β_2	0.121 (0.181)	0.226 (0.152)	0.100 (0.080)
		β_3	1.003 (0.320)	0.884 (0.157)	1.152 (0.153)
		β_4	2.053 (0.162)	1.737 (0.216)	2.053 (0.145)
	600	β_1	0.212 (0.247)	0.335 (0.201)	0.176 (0.129)
		β_2	0.103 (0.082)	0.229 (0.128)	0.096 (0.079)
		β_3	0.998 (0.245)	0.912 (0.165)	1.142 (0.119)
		β_4	2.051 (0.128)	1.773 (0.203)	2.063 (0.120)
	700	β_1	0.166 (0.138)	0.311 (0.186)	0.147 (0.113)
		β_2	0.130 (0.124)	0.222 (0.149)	0.112 (0.083)
		β_3	0.971 (0.213)	0.892 (0.132)	1.107 (0.104)
		β_4	2.056 (0.118)	1.786 (0.187)	2.069 (0.111)
6	500	β_1	0.253 (0.292)	0.350 (0.221)	0.222 (0.162)
		β_2	0.185 (0.258)	0.212 (0.157)	0.150 (0.110)
		β_3	0.303 (0.460)	0.351 (0.170)	0.219 (0.210)
		β_4	0.142 (0.109)	0.200 (0.131)	0.144 (0.102)
		β_5	0.930 (0.302)	0.802 (0.136)	1.144 (0.157)
		β_6	2.003 (0.263)	1.590 (0.148)	2.083 (0.173)
	600	β_1	0.230 (0.205)	0.358 (0.164)	0.193 (0.146)
		β_2	0.159 (0.161)	0.210 (0.107)	0.145 (0.122)
		β_3	0.235 (0.234)	0.353 (0.158)	0.189 (0.165)
		β_4	0.147 (0.105)	0.201 (0.113)	0.132 (0.103)
		β_5	0.913 (0.254)	0.793 (0.048)	1.137 (0.137)
		β_6	2.007 (0.203)	1.589 (0.104)	2.049 (0.143)
	700	β_1	0.190 (0.156)	0.335 (0.165)	0.177 (0.152)
		β_2	0.128 (0.114)	0.183 (0.100)	0.124 (0.113)
		β_3	0.168 (0.138)	0.348 (0.148)	0.163 (0.133)
		β_4	0.126 (0.092)	0.192 (0.107)	0.121 (0.090)
		β_5	0.961 (0.241)	0.805 (0.053)	1.131 (0.126)
		β_6	2.025 (0.150)	1.617 (0.099)	2.052 (0.125)

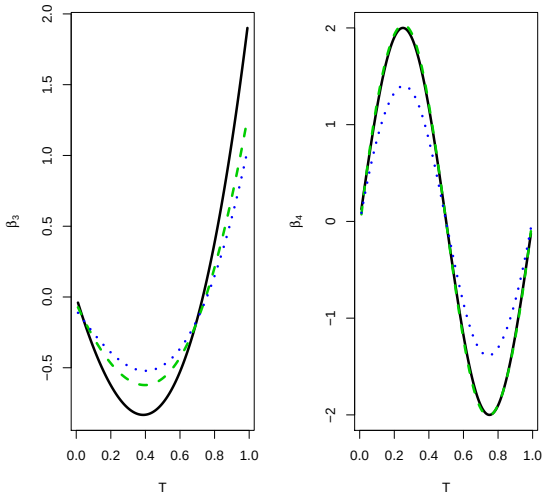

Figure 1. The estimated coefficient functions by the group SCAD (dashed line, green color) and the group Lasso (dotted line, blue color) and true coefficient functions (solid line)

Example 2. We consider the Poisson regression model with the true conditional mean function being

$$E(Y_i|\boldsymbol{x}_i, T_i) \quad = \quad \exp\left(\sum_{j=1}^{p_n} \beta_j(T_i) X_{ij}\right), \quad i = 1, 2, \ldots, n.$$

We set $p_n = 4$ and 6. For $p_n = 4$, let $\beta_1(T_i) = 1, \beta_2(T_i) = 0.5, \beta_3(T_i) = 2T_i(1 - T_i), \beta_4(T_i) = \sin(2\pi T_i)$. For $p_n = 6$, let $\beta_1(T_i) = 1, \beta_2(T_i) = 0.5, \beta_3(T_i) = -1, \beta_4(T_i) = -0.5, \beta_5(T_i) = 2T_i(1-T_i), \beta_6(T_i) = \sin(2\pi T_i)$. The other aspects of the simulation set-up are the same as in Example 1. The simulation results for this example shown in Tables 4-6 demonstrate a similar effect as that of Example 1. Figure 2 shows the estimated varying coefficient functions along with the true function components from the group SCAD and group Lasso. The estimated coefficient functions are given based on the same setting of Figure 1. We can see that $\beta_3(\cdot)$ is under fitting for both the SCAD and Lasso. However, the estimations of $\beta_4(\cdot)$ are very close to the true curve.

Table 4. Number of times each component was selected as a varying coefficient in the 100 replications for the Poisson model.

p_n	n	Method	β_1	β_2	β_3	β_4	β_5	β_6
4	500	SCAD	13	12	85	100		
		Lasso	24	18	86	100		
	600	SCAD	20	13	91	100		
		Lasso	38	24	93	100		
	700	SCAD	21	10	98	100		
		Lasso	43	20	97	100		
6	500	SCAD	33	14	35	13	82	100
		Lasso	33	13	35	8	74	100
	600	SCAD	24	7	24	12	92	100
		Lasso	28	11	26	17	88	100
	700	SCAD	26	9	22	11	96	100
		Lasso	25	12	29	15	94	100

Table 5. Model selection results for the Poisson model

p_n	n	Method	NVC	NTVC
4	500	SCAD	2.10 (0.732)	1.85 (0.359)
		Lasso	2.28 (0.842)	1.86 (0.349)
	600	SCAD	2.24 (0.740)	1.91 (0.288)
		Lasso	2.55 (0.821)	1.93 (0.256)
	700	SCAD	2.29 (0.656)	1.98 (0.141)
		Lasso	2.60 (0.752)	1.97 (0.171)
6	500	SCAD	2.77 (1.118)	1.82 (0.386)
		Lasso	2.63 (1.244)	1.74 (0.441)
	600	SCAD	2.59 (0.805)	1.92 (0.273)
		Lasso	2.70 (1.020)	1.88 (0.327)
	700	SCAD	2.64 (0.847)	1.96 (0.197)
		Lasso	2.75 (0.869)	1.94 (0.239)

Table 6. The root of average mean square error of each component for the Poisson model.

p_n	n		SCAD	Lasso	Oracle
4	500	β_1	0.033 (0.025)	0.041 (0.028)	0.028 (0.022)
		β_2	0.034 (0.029)	0.038 (0.031)	0.034 (0.024)
		β_3	0.168 (0.025)	0.174 (0.028)	0.430 (0.030)
		β_4	0.970 (0.033)	0.949 (0.040)	1.013 (0.036)
	600	β_1	0.029 (0.024)	0.033 (0.025)	0.028 (0.024)
		β_2	0.029 (0.028)	0.034 (0.029)	0.032 (0.026)
		β_3	0.168 (0.021)	0.176 (0.026)	0.425 (0.028)
		β_4	0.972 (0.036)	0.953 (0.037)	1.005 (0.030)
	700	β_1	0.028 (0.020)	0.033 (0.021)	0.031 (0.022)
		β_2	0.028 (0.022)	0.031 (0.025)	0.032 (0.023)
		β_3	0.171 (0.024)	0.178 (0.025)	0.428 (0.027)
		β_4	0.974 (0.029)	0.958 (0.030)	1.003 (0.028)
6	500	β_1	0.025 (0.019)	0.029 (0.022)	0.030 (0.025)
		β_2	0.025 (0.017)	0.027 (0.020)	0.030 (0.025)
		β_3	0.022 (0.018)	0.025 (0.019)	0.030 (0.023)
		β_4	0.021 (0.015)	0.023 (0.016)	0.034 (0.027)
		β_5	0.167 (0.018)	0.168 (0.018)	0.438 (0.046)
		β_6	0.965 (0.037)	0.943 (0.037)	1.004 (0.033)
	600	β_1	0.019 (0.014)	0.022 (0.017)	0.030 (0.024)
		β_2	0.020 (0.016)	0.022 (0.017)	0.033 (0.025)
		β_3	0.017 (0.014)	0.020 (0.015)	0.031 (0.020)
		β_4	0.021 (0.016)	0.022 (0.016)	0.029 (0.022)
		β_5	0.166 (0.012)	0.168 (0.013)	0.432 (0.034)
		β_6	0.972 (0.030)	0.951 (0.030)	1.011 (0.028)
	700	β_1	0.017 (0.014)	0.019 (0.014)	0.025 (0.020)
		β_2	0.018 (0.014)	0.020 (0.014)	0.027 (0.021)
		β_3	0.016 (0.012)	0.019 (0.013)	0.031 (0.020)
		β_4	0.019 (0.014)	0.019 (0.014)	0.028 (0.021)
		β_5	0.166 (0.012)	0.167 (0.012)	0.429 (0.032)
		β_6	0.979 (0.025)	0.959 (0.026)	1.010 (0.031)

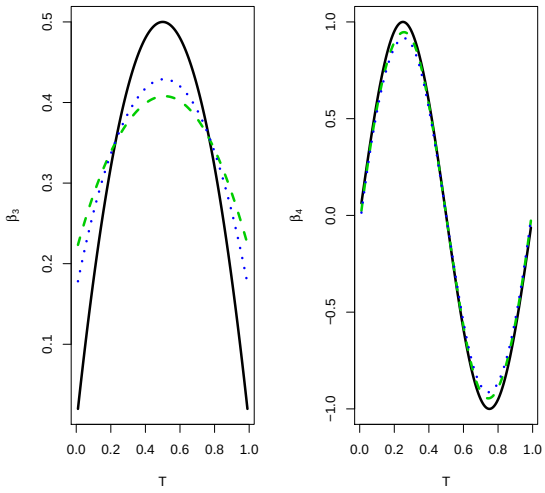

Figure 2. The estimated coefficient functions by the group SCAD (dashed line, green color) and the group Lasso (dotted line, blue color) and true coefficient functions (solid line)

Example 3. We now apply the methodology proposed in this paper to analyze a data set compiled by the General Hospital Burn Center at the University of Southern California. The dataset consists of 981 observations. The binary response variable Y is 1 for those victims who survived their burns and 0 otherwise, the variable T in this application represents age and three covariates were considered including $X_1 =$ sex, $X_2 = \log(\text{burn area } +1)$, binary variable $X_3 =$ oxygen (0 normal, 1 abnormal). We scale the covariate T to [0, 1]. The intercept term is added and the following logistic varying coefficient regression model is considered

$$\text{logit}[\Pr(Y = 1|X_1, X_2, X_3, T)] = \phi_0(T) + \phi_1(T)X_1 + \phi_2(T)X_2 + \phi_3(T)X_3.$$

We are interested in examining whether the regression coefficients vary over different ages. Cubic B-splines with six basis functions are also used to approximate each coefficient. The finial model obtained by the SCAD method is

$$\text{logit}[\Pr(Y = 1|X_1, X_2, X_3, T)] = \phi_0(T) - 0.1268X_1 + \phi_2(T)X_2 - 0.3163X_3.$$

The Lasso method gives the model

$$\text{logit}[\Pr(Y = 1|X_1, X_2, X_3, T)] = \phi_0(T) - 0.1593X_1 - 2.8479X_2 - 0.3633X_3.$$

Cai, Fan and Li (2000)[2] gave the analysis about this data. They obtained that the coefficient functions $\phi_1(T)$ and $\phi_3(T)$ were independent of age. This result is the same as that given by the SCAD in this paper. Figure 3 plots the estimated coefficient functions of ϕ_0 and ϕ_2 from the group SCAD and group Lasso approaches.

To examine the prediction performance of the SCAD and Lasso, we randomly chose 700 observations as training data to fit the model, and the remaining 281 observations are used as test data. This whole process is repeated 100 times. The prediction accuracy of these two methods are shown in Figure 4. From the box plots in Figure 4, we can see that the SCAD method performs better than the Lasso method.

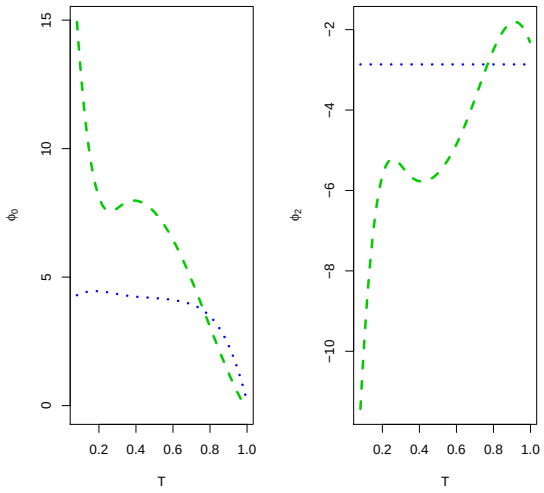

Figure 3. Plots of the estimated coefficient functions of ϕ_0 and ϕ_2 using group SCAD (dashed,green) and group Lasso (dotted, blue).

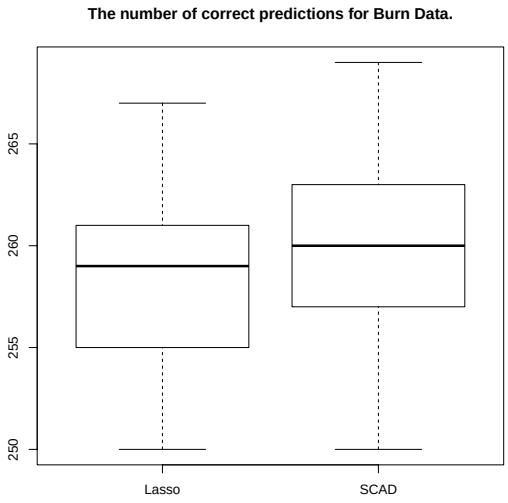

Figure 4. Prediction performance of the SCAD and Lasso based on 100 replicates.

4. Conclusions

In this paper, the structure of the GVCPLM is identified using the group SCAD and it is proved that the varying coefficients and constant coefficients can identify consistently with probability tending to one under certain regularity conditions. Furthermore, the convergence rate of the proposed estimator is obtained and the model selection is considered only.

Acknowledgments

Mingqiu Wang's research was supported by the National Natural Science Foundation of China (11401340) and Doctor Scientific Research Foundation, Qufu Normal University (bsqd2012041). Xiuli Wang's research was supported by the Foundation of Qufu Normal University (xkj201518).

References

[1] Ahmad, I, Leelahanon, S, Li, Q. *Efficient estimation of a semiparametric partially linear varying coefficient model.* Annals of Statistics. **33**, 258–283, 2005.

[2] Cai, Z, Fan, J, Li, R. *Efficient estimation and inferences for varying-coefficient models.* Journal of the American Statistical Association. **95**, 888–902, 2000.

[3] Fan, J, Li, R. *Variable selection via nonconcave penalized likelihood and its oracle properties.* Journal of the American Statistical Association. **6**, 1348–1360, 2001.

[4] Hastie, T, Tibshirani, R. *Varying-coefficient models.* Journal of the Royal Statistical Society, Series B. **55**, 757–796, 1993.

[5] Hu, T, Cui, H. *Robust estimates in generalised varying-coefficient partially linear models.* Journal of Nonparametric Statistics. **22**, 737–754, 2010.

[6] Hu, T, Xia, Y. *Adaptive semi-varying coefficient model selection.* Statistica Sinica. **22**, 575–599, 2012.

[7] Huang, J, Horowitz, JL, Wei FR. *Variable selection in nonparametric additive models.* Annals of Statistics. **38**, 2282–2313, 2010.

[8] Huang, J, Wei, F, Ma, S. *Semiparametric regression pursuit.* Statistica Sinica. **22**, 1403–1426, 2012.

[9] Huang, JHZ, Wu, CO, Zhou, L. *Polynomial spline estimation and inference for varying coefficient models with longitudinal data.* Statistica Sinica. **14**, 763–788, 2004.

[10] Kim, Y, Choi, H, Oh, H. *Smoothly clipped absolute deviation on high dimensions.* Journal of the American Statistical Association. **103**, 1656–1673, 2008.

[11] Kai, B, Li, R, Zou, H. *New efficient estimation and variable methods for semiparametric varying-coefficient partially linear models.* Annals of Statistics. **39**, 305–332, 2011.

[12] Lam, C, Fan, J. *Profile-kernel likelihood inference with diverging number of parameters.* Annals of Statistics. **36**, 2232–2260, 2008.

[13] Li, Q, Huang, CJ, Li, D, Fu, TT. *Semiparametric smooth coefficient models.* Journal of Business and Economic Statistics. **20**, 412–422, 2002.

[14] Li, R, Liang, H. *Variable selection in semiparametric regression modeling.* Annals of Statistics. **36**, 261–286, 2008.

[15] Li G, Xue L, Lian H. *SCAD-penalised generalised additive models with non-polynomial dimensionality.* Journal of Nonparametric Statistics, **24**, 681–697, 2012.

[16] Li, G, Lin, L, Zhu, L. *Empirical likelihood for a varying coefficient partially linear model with diverging number of parameters.* Journal of Multivariate Analysis. **105**, 85–111, 2012.

[17] Lian, H. *Variable selection for high-dimensional generalized varying-coefficient models.* Statistica Sinica. **22**, 1563–1588, 2012.

[18] Lian, H, Chen, X, Yang, JY. *Identification of partially linear structure in additive models with an application to gene expression prediction from sequences.* Biometrics. **68**, 437–445, 2012.

[19] Lian, H, Du, P, Li, Y, Liang, H. *Partially linear structure identification in generalized additive models with NP-dimensionality.* Computational Statistics & Data Analysis. **80**, 197–208, 2014.

[20] Lian, H, Liang, H, Ruppert, D. *Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models.* Statistica Sinica. **25**, 591–607, 2015.

[21] Lu, Y:. *Generalized partially linear varying-coefficient models.* Journal of Statistical Planning and Inference. **138**, 901–914, 2008.

[22] McCullagh, P, Nelder, JA. *Generalized Linear Models.* Chapman and Hall, London (1989).

[23] Schwarz, G. *Estimating the dimension of a model.* Annals of Statistics. **6**, 461–464, 1978.

[24] Schumaker, LL. *Spline Functions: Basic Theory.* Wiley, New York (1981).

[25] Tang, Y, Wang, HJ, Zhu, Z, Song, X. *A unified variable selection approach for varying coefficient models.* Statistica Sinica. **22**, 601–628, 2012.

[26] Wang, D, Kulasekera, KB. *Parametric component detection and variable selection in varying-coefficient partially linear models.* Journal of Multivariate Analysis. **112**, 117–129, 2012.

[27] Wang, H., Li, R., Tsai, C.L. *Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method.* Biometrika, **94**, 553–568, 2007.

[28] Wang, M, Song, L. *Identification for semiparametric varying coefficient partially linear models. Statist.* Statistics & Probability Letters. **83**, 1311–1320, 2013.

[29] Xia, Y, Zhang, W, Tong, H. *Efficient estimation for semivarying-coefficient models.* Biometrika. **91**, 661–681, 2004.

[30] Zhang, H, Cheng, G, Liu, Y. *Linear or nonlinear? Automatic structure discovery for partially linear models.* Journal of the American Statistical Association. **106**, 1099–1112, 2011.

[31] Zhou ,S, Shen, X, Wolfe, DA. *Local asymptotics for regression splines and confidence regions.* Annals of Statistics. **26**, 1760–1782, 1998.