Journal of Machine Learning Research 13 (2012) 1973-1998 Submitted 3/11; Revised 1/12; Published 6/12

Variable Selection in High-dimensional Varying-coefficient Models with Global Optimality

Lan Xue XUEL@STAT.OREGONSTATE.EDU
Department of Statistics
Oregon State University
Corvallis, OR 97331-4606, USA

Annie Qu ANNIEQU@ILLINOIS.EDU
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL 61820-3633, USA

Editor: Xiaotong Shen

Abstract

The varying-coefficient model is flexible and powerful for modeling the dynamic changes of regression coefficients. It is important to identify significant covariates associated with response variables, especially for high-dimensional settings where the number of covariates can be larger than the sample size. We consider model selection in the high-dimensional setting and adopt difference convex programming to approximate the L_0 penalty, and we investigate the global optimality properties of the varying-coefficient estimator. The challenge of the variable selection problem here is that the dimension of the nonparametric form for the varying-coefficient modeling could be infinite, in addition to dealing with the high-dimensional linear covariates. We show that the proposed varying-coefficient estimator is consistent, enjoys the oracle property and achieves an optimal convergence rate for the non-zero nonparametric components for high-dimensional data. Our simulations and numerical examples indicate that the difference convex algorithm is efficient using the coordinate decent algorithm, and is able to select the true model at a higher frequency than the least absolute shrinkage and selection operator (LASSO), the adaptive LASSO and the smoothly clipped absolute deviation (SCAD) approaches.

Keywords: coordinate decent algorithm, difference convex programming, L_0- regularization, large-p small-n, model selection, nonparametric function, oracle property, truncated L_1 penalty

1. Introduction

High-dimensional data occur very frequently and are especially common in biomedical studies including genome studies, cancer research and clinical trials, where one of the important scientific interests is in dynamic changes of gene expression, long-term effects for treatment, or the progression of certain diseases.

We are particularly interested in the varying-coefficient model (Hastie and Tibshirani, 1993; Ramsay and Silverman, 1997; Hoover et al., 1998; Fan and Zhang, 2000; Wu and Chiang, 2000; Huang, Wu and Zhou, 2002, 2004; Qu and Li, 2006; Fan and Huang, 2005; among others) as it is powerful for modeling the dynamic changes of regression coefficients. Here the response variables are associated with the covariates through linear regression, but the regression coefficients can vary and are modeled as a nonparametric function of other predictors.

In the case where some of the predictor variables are redundant, the varying-coefficient model might not be able to produce an accurate and efficient estimator. Model selection for significant predictors is especially critical when the dimension of covariates is high and possibly exceeds the sample size, but the number of nonzero varying-coefficient components is relatively small. This is because even a single predictor in the varying-coefficient model could be associated with a large number of unknown parameters involved in the nonparametric functions. Inclusion of high-dimensional redundant variables can hinder efficient estimation and inference for the non-zero coefficients.

Recent developments in variable selection for varying-coefficient models include Wang, Li and Huang (2008) and Wang and Xia (2009), where the dimension of candidate models is finite and smaller than the sample size. Wang, Li and Huang (2008) considered the varying-coefficient model in a longitudinal data setting built on the SCAD approach (Fan and Li, 2001; Fan and Peng, 2004), and Wang and Xia (2009) proposed the use of local polynomial regression with an adaptive LASSO penalty. For the high-dimensional case when the dimension of covariates is much larger than the sample size, Wei, Huang and Li (2011) proposed an adaptive group LASSO approach using B-spline basis approximation. The SCAD penalty approach has the advantages of unbiasedness, sparsity and continuity. However, the SCAD approach involves non-convex optimization through local linear or quadratic approximations (Hunter and Li, 2005; Zou and Li, 2008), which is quite sensitive to the initial estimator. In general, the global minimum is not easily obtained for non-convex function optimization. Kim, Choi and Oh (2008) have improved SCAD model selection using the difference convex (DC) algorithm (An and Tao, 1997; Shen et al., 2003). Still, the existence of global optimality for the SCAD has not been investigated for the case that the dimension of covariates exceeds the sample size. Alternatively, the adaptive LASSO and the adaptive group LASSO approaches are easier to implement due to solving the convex optimization problem. However, the adaptive LASSO algorithm requires the initial estimators to be consistent, and such a requirement could be difficult to obtain in high-dimensional settings.

Indeed, obtaining consistent initial estimators of the regression parameters is more difficult than the model selection problem when the dimension of covariates exceeds the sample size, since if the initial estimator is already close to the true value, then performing model selection is much less challenging. So far, most model selection algorithms rely on consistent LASSO estimators as initial values. However, the irrepresentable assumption (Zhao and Yu, 2006) to obtain consistent LASSO estimators for high-dimensional data is unlikely to be satisfied, since most of the covariates are correlated. When the initial consistent estimators are no longer available, the adaptive LASSO and the SCAD algorithm based on either local linear or quadratic approximations are likely to fail.

To overcome the aforementioned problems, we approximate the L_0 penalty effectively as the L_0 penalty is considered to be optimal for achieving sparsity and unbiasedness, and is optimal even for the high-dimensional data case. However, the challenge of L_0 regularization is computational difficulty due to its non-convexity and non-continuity. We use a newly developed truncated L_1 penalty (TLP, Shen, Pan and Zhu, 2012) for the varying-coefficient model which is piecewise linear and continuous to approximate the non-convex penalty function. The new method intends to overcome the computational difficulty of the L_0 penalty while preserving the optimality of the L_0 penalty. The key idea is to decompose the non-convex penalty function by taking the difference between two convex functions, thereby transforming a non-convex problem into a convex optimization problem.

One of the main advantages of the proposed approach is that the minimization process does not depend on the initial estimator, which could be hard to obtain when the dimension of covariates is high. In addition, the proposed algorithm for the varying-coefficient model is computationally effi-

cient. This is reflected in that the proposed model selection performs better than existing approaches such as SCAD in the high-dimensional case, based on our simulation and as applied to HIV AIDs data, with a much higher frequency of choosing the correct model. The improvement is especially significant when the dimension of covariates is much higher than the sample size.

We derive model selection consistency for the proposed method and show that it possesses the oracle property when the dimension of covariates exceeds the sample size. Note that the theoretical derivation of asymptotic properties and global optimality results are rather challenging for varying-coefficient model selection, as we are dealing with an infinite dimension of the nonparametric component in addition to the high-dimensional covariates. In addition, the optimal rate of convergence for the non-zero nonparametric components can be achieved in high-dimensional varying-coefficient models. The theoretical techniques applied in this project are innovative as there is no existing theoretical result on global optimality for high-dimensional model selection in the varying-coefficient model framework.

The paper is organized as follows. Section 2 provides the background of varying-coefficient models. Section 3 introduces the penalized polynomial spline procedure for selecting varying-coefficient models when the dimension of covariates is high, provides the theoretical properties for model selection consistency and establishes the relationship between the oracle estimator and the global and local minimizers. Section 4 provides tuning parameter selection, and the coordinate decent algorithm for model selection implementation. Section 5 demonstrates simulations and a data example for high-dimensional data. The last section provides concluding remarks and discussion.

2. Varying-coefficient Model

Let $(\mathbf{X}_i, U_i, Y_i), i = 1, \ldots, n$, be random vectors that are independently and identically distributed as $(\mathbf{X}, U, Y)$, where $\mathbf{X} = (X_1, \ldots, X_d)^T$ and a scalar U are predictor variables, and Y is a response variable. The varying-coefficient model (Hastie and Tibshirani, 1993) has the following form:

$$Y_i = \sum_{j=1}^{d} \beta_j(U_i) X_{ij} + \varepsilon_i, \tag{1}$$

where X_{ij} is the jth component of $\mathbf{X}_i$, $\beta_j(\cdot)$'s are unknown varying-coefficient functions, and ε_i is a random noise with mean 0 and finite variance σ^2. The varying-coefficient model is flexible in that the responses are linearly associated with a set of covariates, but their regression coefficients can vary with another variable U. We will call U the index variable and $\mathbf{X}$ the linear covariates. In practice, some of the linear covariates may be irrelevant to the response variable, with the corresponding varying-coefficient functions being zero almost surely. The goal of this paper is to identify the irrelevant linear covariates and estimate the nonzero coefficient functions for the relevant ones.

In many applications, such as microarray studies, the total number of the available covariates d can be much larger than the sample size n, although we assume that the number of relevant ones is fixed. In this paper, we propose a penalized polynomial spline procedure in variable selection for the varying-coefficient model where the number of linear covariates d is much larger than n. The proposed method is easy to implement and fast to compute. In the following, without loss of generality, we assume there exists an integer d_0 such that $0 < E\left[\beta_j^2(U)\right] < \infty$ for $j = 1, \ldots, d_0$, and $E\left[\beta_j^2(U)\right] = 0$ for $j = d_0, \ldots, d$. Furthermore, we assume that only the first d_0 covariates in $\mathbf{X}$ are relevant, and that the rest of the covariates are redundant.

3. Model Selection in High-dimensional Data

In our estimation procedure, we first approximate the smooth functions $\{\beta_j(\cdot)\}_{j=1}^d$ in (1) by polynomial splines. Suppose U takes values in $[a,b]$ with $a<b$. Let υ_j be a partition of the interval $[a,b]$, with N_n interior knots

$$\upsilon_j = \{a=\upsilon_{j,0}<\upsilon_{j,1}<\cdots<\upsilon_{j,N_n}<\upsilon_{j,N_n+1}=b\}.$$

Using υ_j as knots, the polynomial splines of order $p+1$ are functions which are p-degree (or less) of polynomials on intervals $[\upsilon_{j,i},\upsilon_{j,i+1}), i=0,\ldots,N_n-1$, and $[\upsilon_{j,N_n},\upsilon_{j,N_n+1}]$, and have $p-1$ continuous derivatives globally. We denote the space of such spline functions by φ_j. The advantage of polynomial splines is that they often provide good approximations of smooth functions with only a small number of knots.

Let $\{B_{jl}(\cdot)\}_{l=1}^{J_n}$ be a set of B-spline bases of φ_j with $J_n=N_n+p+1$. Then for $j=1,\ldots,d$,

$$\beta_j(\cdot)\approx s_j(\cdot)=\sum_{l=1}^{J_n}\gamma_{jl}B_{jl}(\cdot)=\gamma_j^T\mathbf{B}_j(\cdot),$$

where $\gamma_j=(\gamma_{j1},\ldots,\gamma_{jJ_n})^T$ is a set of coefficients, and $\mathbf{B}_j(\cdot)=(B_{j1}(\cdot),\ldots,B_{jJ_n}(\cdot))^T$ are B- spline bases. The standard polynomial spline method (Huang, Wu and Zhou, 2002) estimates the coefficient functions $\{\beta_j(\cdot)\}_{j=1}^d$ by spline functions which minimize the sum of squares

$$\left(\widetilde{\beta}_1,\ldots,\widetilde{\beta}_d\right)=\underset{s_j\in\varphi_j,j=1,\ldots,d}{\operatorname{argmin}}\frac{1}{2n}\sum_{i=1}^n\left[Y_i-\sum_{j=1}^d s_j(U_i)X_{ij}\right]^2.$$

Equivalently, in terms of B-spline basis, it estimates $\gamma=\left(\gamma_1^T,\ldots,\gamma_d^T\right)^T$ by

$$\widetilde{\gamma}=\left(\widetilde{\gamma}_1^T,\ldots,\widetilde{\gamma}_d^T\right)^T=\underset{\gamma_j,j=1,\ldots,d}{\operatorname{argmin}}\frac{1}{2n}\sum_{i=1}^n\left[Y_i-\sum_{j=1}^d\gamma_j^T\mathbf{Z}_{ij}\right]^2, \tag{2}$$

where $\mathbf{Z}_{ij}=\mathbf{B}_j(U_i)X_{ij}=(B_{j1}(U_i)X_{ij},\ldots,B_{jJ_n}(U_i)X_{ij})^T$. However, the standard polynomial spline approach fails to reduce model complexity when some of the linear covariates are redundant, and furthermore is not able to obtain parameter estimation when the dimension of model d is larger than the sample size n. Therefore, to perform simultaneous variable selection and model estimation, we propose minimizing the penalized sum of squares

$$L_n(\mathbf{s})=\frac{1}{2n}\sum_{i=1}^n\left[Y_i-\sum_{j=1}^d s_j(U_i)X_{ij}\right]^2+\lambda_n\sum_{j=1}^d p_n\left(\|s_j\|_n\right), \tag{3}$$

where $\mathbf{s}=\mathbf{s}(\cdot)=(s_1(\cdot),\ldots,s_d(\cdot))^T$, and $\|s_j\|_n=\left(\sum_{i=1}^n s_j^2(U_i)X_{ij}^2/n\right)^{1/2}$ is the empirical norm. In terms of the B-spline basis, (3) is equivalent to

$$L_n(\gamma)=\frac{1}{2n}\sum_{i=1}^n\left[Y_i-\sum_{j=1}^d\gamma_j^T\mathbf{Z}_{ij}\right]^2+\lambda_n\sum_{j=1}^d p_n\left(\|\gamma_j\|_{W_j}\right), \tag{4}$$

where $\|\gamma_j\|_{W_j} = \sqrt{\gamma_j^T \mathbf{W}_j \gamma_j}$ with $\mathbf{W}_j = \sum_{i=1}^{n} \mathbf{Z}_{ij}\mathbf{Z}_{ij}^T/n$. The formulation (3) is quite general. In particular, for a linear model with $\beta_j(u) = \beta_j$ and the linear covariates being standardized with $\sum_{i=1}^n X_{ij}/n = 0$ and $\sum_{i=1}^n X_{ij}^2/n = 1$ for $j = 1,\ldots,d$, (3) reduces to a family of variable selection methods for linear models with the penalty $p_n\left(\|s_j\|_n\right) = p_n\left(|\beta_j|\right)$. For instance, the L_1 penalty $p_n(|\beta|) = |\beta|$ results in LASSO (Tibshirani, 1996), and the smoothly clipped absolute deviation penalty results in SCAD (Fan and Li, 2001). In this paper, we consider a rather different approach for the penalty function such that

$$p_n(\beta) = p(\beta, \tau_n) = \min(|\beta|/\tau_n, 1), \tag{5}$$

which is called a truncated L_1-penalty (TLP) function, as proposed in Shen, Pan and Zhu (2012). In (5), the additional tuning parameter τ_n is a threshold parameter determining which individual components are to be shrunk towards to zero, or not. As pointed out by Shen, Pan and Zhu (2012), the TLP corrects the bias of the LASSO induced by the convex L_1-penalty and also reduces the computational instability of the L_0-penalty. The TLP is able to overcome the computation difficulty for solving non-convex optimization problems by applying difference convex programming, which transforms non-convex problems into convex optimization problems. This leads to significant computational advantages over its smooth counterparts, such as the SCAD (Fan and Li, 2001) and the minimum concavity penalty (MCP, Zhang, 2010). In addition, the TLP works particularly well for high-dimensional linear regression models as it does not depend on initial consistent estimators of coefficients, which could be difficult to obtain when d is much larger than n. In this paper, we will investigate the local and global optimality of the TLP for variable selection in varying-coefficient models in the high-dimensional case when $d \gg n$, and n goes to infinity.

Here we obtain $\widehat{\gamma}$ by minimizing $L_n(\gamma)$ in (4). As a result, for any $u \in [a,b]$, the estimators of the unknown varying-coefficient functions in (1) are given as

$$\widehat{\beta}_j(u) = \sum_{l=1}^{J_n} \widehat{\gamma}_{jl} B_{jl}(u), \ j = 1,\ldots,d. \tag{6}$$

Let $\widetilde{\gamma}^{(o)} = (\widetilde{\gamma}_1,\ldots,\widetilde{\gamma}_{d_0},0,\ldots,0)^T$ be the oracle estimator with the first d_0 elements being the standard polynomial estimator (2) of the true model consisting of only the first d_0 covariates. The following theorems establish the asymptotic properties of the proposed estimator. We only state the main results here and relegate the regularity conditions and proofs to the Appendix.

Theorem 1 *Let $A_n(\lambda_n, \tau_n)$ be the set of local minima of (4). Under conditions (C1-C7) in the Appendix, the oracle estimator is a local minimizer with probability tending to 1, that is,*

$$P\left(\widetilde{\gamma}^{(o)} \in A_n(\lambda_n, \tau_n)\right) \to 1,$$

as $n \to \infty$.

Theorem 2 *Let $\widehat{\gamma} = (\widehat{\gamma}_1,\ldots,\widehat{\gamma}_d)^T$ be the global minima of (4). Under conditions (C1-C6), (C8) and (C9) in the Appendix, the estimator by minimizing (4) enjoys the oracle property, that is,*

$$P\left(\widehat{\gamma} = \widetilde{\gamma}^{(o)}\right) \to 1,$$

as $n \to \infty$.

Theorem 1 guarantees that the oracle estimator must fall into the local minima set. Theorem 2, in addition, provides sufficient conditions such that the global minimizer by solving the non-convex objective function in (4) is also the oracle estimator.

In addition to the results of model selection consistency, we also establish the oracle property for the non-zero components of the varying-coefficients. For any $u\in[a,b]$, let $\widehat{\beta}^{(1)}(u) = \left(\widehat{\beta}_1(u),\ldots,\widehat{\beta}_{d_0}(u)\right)^T$ be the estimator of the first d_0 varying-coefficient functions which are non-zero and are defined in (6) with $\widehat{\gamma}$ being the global minima of (4). Theorem 3 establishes the asymptotic normality of $\widehat{\beta}^{(1)}(u)$ with the optimal rate of convergence.

Theorem 3 *Under conditions (C1) - (C6), (C8) and (C9) given in the Appendix, and if* $\lim N_n \log N_n/n = 0$*, then for any* $u \in [a,b]$,

$$\left\{V\left(\widehat{\beta}^{(1)}(u)\right)\right\}^{-1/2}\left(\widehat{\beta}^{(1)}(u)-\beta_0^{(1)}(u)\right)\to N(0,\mathbf{I})$$

in distribution, where $\beta_0^{(1)}(u) = (\beta_{01}(u),\ldots,\beta_{0d_0}(u))^T$, $\mathbf{I}$ *is a* $d_0\times d_0$ *identity matrix, and*

$$V\left(\widehat{\beta}^{(1)}(u)\right) = \mathbf{B}^{(1)}(u)\left(\sum_{i=1}^n \mathbf{A}_i^{(1)T}\mathbf{A}_i^{(1)}\right)^{-1}\mathbf{B}^{(1)}(u) = O_p(N_n/n),$$

in which $\mathbf{B}^{(1)}(u) = \left(\mathbf{B}_1^T(u),\ldots,\mathbf{B}_{d_0}^T(u)\right)^T$, *and* $\mathbf{A}_i^{(1)} = \left(\mathbf{B}_1^T(U_i)X_{i1},\ldots,\mathbf{B}_{d_0}^T(U_i)X_{id_0}\right)^T$ *with* $\mathbf{B}_j^T(U_i)X_{ij} = (B_{j1}(U_i)X_{ij},\ldots,B_{jJ_n}(U_i)X_{ij})$.

4. Implementation

In this section, we extend the difference convex (DC) algorithm of Shen, Pan and Zhu (2012) to solve the nonconvex minimization in (4) for varying-coefficient models. In addition, we provide the tuning parameter selection criteria.

4.1 An Algorithm

The idea of the DC algorithm is to decompose a non-convex object function into a difference between two convex functions. Then the final solution is obtained iteratively by minimizing a sequence of upper convex approximations of the non-convex objective function. Specifically, we decompose the penalty in (5) as $p_n(\beta) = p_{n1}(\beta) - p_{n2}(\beta)$, where $p_{n1}(\beta) = |\beta|/\tau_n$ and $p_{n2}(\beta) = \max(|\beta|/\tau_n - 1, 0)$. Note that both $p_{n1}(\cdot)$ and $p_{n2}(\cdot)$ are convex functions. Therefore, we can decompose the non-convex objective function $L_n(\gamma)$ in (4) as a difference between two convex functions,

$$L_n(\gamma) = L_{n1}(\gamma) - L_{n2}(\gamma),$$

where

$$\begin{aligned} L_{n1}(\gamma) &= \frac{1}{2n}\sum_{i=1}^n\left[Y_i - \sum_{j=1}^d \gamma_j^T\mathbf{Z}_{ij}\right]^2 + \lambda_n\sum_{j=1}^d p_{n1}\left(\|\gamma_j\|_{W_j}\right), \\ L_{n2}(\gamma) &= \lambda_n\sum_{j=1}^d p_{n2}\left(\|\gamma_j\|_{W_j}\right). \end{aligned}$$

Let $\widehat{\gamma}^{(0)}$ be an initial value. From our experience, the proposed algorithm does not rely on initial consistent estimators of coefficients so we have used $\widehat{\gamma}^{(0)} = \mathbf{0}$ in the implementations. At iteration m, we set $L_n^{(m)}(\gamma)$, an upper approximation of $L_n(\gamma)$, equal to

$$\begin{aligned} & L_{n1}(\gamma) - \left[L_{n2}\left(\widehat{\gamma}^{(m-1)}\right) + \lambda_n \sum_{j=1}^{d} \left(\|\gamma_j\|_{W_j} - \left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} \right) p_{n2}^{'}\left(\left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} \right) \right] \\ \approx & \frac{1}{2n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{d} \gamma_j^T \mathbf{Z}_{ij} \right]^2 + \frac{\lambda_n}{\tau_n}\sum_{j=1}^{d} \|\gamma_j\|_{W_j} I\left(\left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} \le \tau_n \right) \\ & -L_{n2}\left(\widehat{\gamma}^{(m-1)}\right) + \frac{\lambda_n}{\tau_n}\sum_{j=1}^{d} \left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} I\left(\left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} > \tau_n \right), \end{aligned}$$

where $p_{n2}^{'}\left(\left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} \right) = \frac{1}{\tau_n} I(\left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} > \tau_n)$ is the subgradient of p_{n2}**.** Since the last two terms of the above equation do not depend on γ, therefore at iteration m,

$$\widehat{\gamma}^{(m)} = \underset{\gamma_j, j=1,\ldots,d}{\operatorname{argmin}} \left\{ \frac{1}{2n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{d} \gamma_j^T \mathbf{Z}_{ij} \right]^2 + \sum_{j=1}^{d} \lambda_{nj} \|\gamma_j\|_{W_j} \right\}, \tag{7}$$

where $\lambda_{nj} = \frac{\lambda_n}{\tau_n} I\left(\left\|\widehat{\gamma}_j^{(m-1)}\right\|_{W_j} \le \tau_n \right)$. Then it reduces to a group lasso with component-specific tuning parameter λ_{nj}. It can be solved by applying the coordinate-wise descent (CWD) algorithm as in Yuan and Lin (2006). To be more specific, let $\mathbf{Z}_{ij}^* = \mathbf{W}_j^{-1/2}\mathbf{Z}_{ij}$ and $\gamma_j^* = \mathbf{W}_j^{1/2}\gamma_j$. Then the minimization problem in (7) reduces to

$$\widehat{\gamma}^{*(m)} = \underset{\gamma_j^*, j=1,\ldots,d}{\operatorname{argmin}} \left\{ \frac{1}{2n}\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{d} \gamma_j^{*T} \mathbf{Z}_{ij}^* \right]^2 + \lambda_{nj}\sum_{j=1}^{d} \|\gamma_j^*\|_2 \right\}. \tag{8}$$

Then the CWD algorithm minimizes (8) in each component while fixing the remaining components at their current value. For the jth component, $\widehat{\gamma}_j^{*(m)}$ is updated by

$$\gamma_j^{*(m)} = \left(1 - \frac{\lambda_{nj}}{\|S_j\|_2} \right)_+ S_j, \tag{9}$$

where $S_j = \mathbf{Z}_j^{*T}\left(\mathbf{Y} - \mathbf{Z}^* \gamma_{-j}^{*(m)} \right)$ with $\gamma_{-j}^{*(m)} = \left(\gamma_1^{*(m)T}, \ldots, \gamma_{j-1}^{*(m)T}, \mathbf{0}^T, \gamma_{j+1}^{*(m)T}, \ldots, \gamma_d^{*(m)T} \right)^T$, $\mathbf{Z}_j^* = \left(\mathbf{Z}_{1j}^*, \ldots, \mathbf{Z}_{nj}^* \right)^T$, $\mathbf{Z}^* = \left(\mathbf{Z}_1^*, \ldots, \mathbf{Z}_d^* \right)$ and $(x)_+ = xI_{\{x \ge 0\}}$. The solution to (8) can therefore be obtained by iteratively applying Equation (9) to $j = 1, \ldots, d$ until convergence.

The above algorithm is piece-wise linear and therefore it is computationally efficient. The penalty part in (7) only involves a large L_2-norm of the varying-coefficient function, implying that there is no shrinkage for the non-zero components with a large magnitude of coefficients. In addition, the above algorithm can capture weak signals of varying-coefficients, and meanwhile is able to

obtain the sparsest solution through tuning the additional thresholding parameters τ_n. The involvement of the additional tuning of τ_n makes the TLP a flexible optimization procedure.

The minimization in (4) can achieve the global minima if the leading convex function can be approximated, and it is called the outer approximation method (Breiman and Cutler, 1993). However, it has a slower convergence rate. Here we approximate the trailing convex function with fast computation, and it leads to a good local minimum if it is not global (Shen, Pan and Zhu, 2012). It can achieve the global minimizer if it is combined with the branch-and-bound method (Liu, Shen and Wong, 2005), which searches through all the local minima with an additional cost in computation. This contrasts to the SCAD or adaptive LASSO approaches which are based on local approximation. Achieving the global minimum is particularly important if the dimension of covariates is high, as the number of possible local minima increases dramatically as p increases. Therefore, any local approximation algorithm which relies on initial values likely fails.

4.2 Tuning Parameter Selection

The performance of the proposed spline TLP method crucially depends on the choice of tuning parameters. One needs to choose the knot sequences in the polynomial spline approximation and λ_n, τ_n in the penalty function. For computation convenience, we use equally spaced knots with the number of interior knots $N_n = [n^{1/(2p+3)}]$, and select only λ_n, τ_n. A similar strategy for knot selection can also be found in Huang, Wu and Zhou (2004), and Xue, Qu and Zhou (2010). Let $\theta_n = (\lambda_n, \tau_n)$ be the parameters to be selected. For faster computation, we use K-fold cross-validation to select θ_n, with $K = 5$ in the implementation. The full data T is randomly partitioned into K groups of about the same size, denoted as T_v , for $v = 1, \ldots, K$. Then for each v, the data $T - T_v$ is used for estimation and T_v is used for validation. For any given θ_n, let $\hat{\beta}_j^{(v)}(\cdot, \theta_n)$ be the estimators of $\beta_j(\cdot)$ using the training data $T - T_v$ for $j = 1, \ldots, d$. Then the cross-validation criterion is given as

$$\mathrm{CV}(\theta_n) = \sum_{v=1}^{K} \sum_{i \in T_v} \left\{ Y_i - \sum_{j=1}^{d} \hat{\beta}_j^{(v)}(U_i, \theta_n) X_{ij} \right\}^2 .$$

We select $\hat{\theta}_n$ by minimizing $\mathrm{CV}(\theta_n)$.

5. Simulation and Application

In this section, we conduct simulation studies to demonstrate the finite sample performance of the proposed method. We also illustrate the proposed method with an analysis of an AIDS data set. The total average integrated squared error (TAISE) is evaluated to assess estimation accuracy. Let $\widehat{\beta}^{(r)}$ be the estimator of a nonparametric function β in the r-th ($1 \le r \le R$) replication and $\{u_m\}_{m=1}^{n_{\mathrm{grid}}}$ be the grid points where $\widehat{\beta}^{(r)}$ is evaluated. We define $\mathrm{AISE}\left(\widehat{\beta}\right) = \frac{1}{R}\sum_{r=1}^{R} \frac{1}{n_{\mathrm{grid}}} \sum_{m=1}^{n_{\mathrm{grid}}} \left\{ \beta(u_m) - \widehat{\beta}^{(r)}(u_m) \right\}^2$, and $\mathrm{TAISE} = \sum_{l=1}^{d} \mathrm{AISE}\left(\widehat{\beta}_l\right)$. Let $\mathcal{S}$ and $\mathcal{S}_0$ be the selected and true index sets containing significant variables, respectively. We say $\mathcal{S}$ is correct if $\mathcal{S} = \mathcal{S}_0$; $\mathcal{S}$ overfits if $\mathcal{S}_0 \subset \mathcal{S}$ but $\mathcal{S}_0 \neq \mathcal{S}$; and $\mathcal{S}$ underfits if $\mathcal{S}_0 \not\subset \mathcal{S}$. In all simulation studies, the total number of simulations is 500.

5.1 Simulated Example

We consider the following varying-coefficient model

$$Y_i = \sum_{j=1}^{d} \beta_j(U_i) X_{ij} + \varepsilon_i,\ i = 1, \ldots, 200, \tag{10}$$

where the index variables U_i are generated from a Uniform $[0,1]$, and the linear covariates $\mathbf{X}_i$ are generated from a multivariate normal distribution with mean $\mathbf{0}$ and $Cov(X_{ij}, X_{ij'}) = 0.5^{|j-j'|}$, the noises ε_i are generated from a standard normal distribution, and the coefficient functions are of the forms

$$\beta_1(u) = \sin(2\pi u),\ \ \beta_2(u) = (2u-1)^2 + 0.5,\ \ \beta_3(u) = \exp(2u-1) - 1,$$

and $\beta_j(u) = 0$ for $j = 4, \ldots, d$. Therefore only the first three covariates are relevant for predicting the response variable, and the rest are null variables and do not contribute to the model prediction. We consider the model (10) with $d = 10$, 100, 200, or 400 to examine the performance of model selection and estimation when d is smaller than, close to, or exceeds the sample size.

We apply the proposed varying-coefficient TLP with a linear spline. The simulation results based on the cubic spline are not provided here as they are quite similar to those based on the linear spine. The tuning parameters are selected using the five-fold cross-validation procedure as described in Section 4.2. We compare the TLP approach to a penalized spline procedure with the SCAD penalty, the group LASSO (LASSO) and the group adaptive LASSO (AdLASSO) as described in Wei, Huang and Li (2011). For the SCAD penalty, the first order derivative of $p_n(\cdot)$ in (4) is given as $p_n'(\theta) = I(\theta \le \lambda_n) + \frac{(a\lambda_n - \theta)_+}{(a-1)\lambda_n} I(\theta > \lambda_n)$, and we set $a = 3.7$ as in Fan and Li (2001). For all procedures, we select the tuning parameters using a five-fold cross-validation procedure for fair comparison. To assess the estimation accuracy of the penalized methods, we also consider the standard polynomial spline estimations of the oracle model (ORACLE). The oracle model only contains the first three relevant variables and is only available in simulation studies where the true information is known.

Table 1 summarizes the simulation results. It gives the relative TAISEs (RTAISE) of the penalized spline methods (TLP, SCAD, LASSO, AdLASSO) to the ORACLE estimator. It also reports the percentage of correct fitting(C), underfitting(U) and overfitting(O) over 200 simulation runs for the penalized methods. When $d = 10$, the performance of the TLP, SCAD, LASSO and AdLASSO are comparable, with TLP being slightly better the rest. But as the dimension d increases, Table 1 clearly shows that the TLP outperforms the other procedures. The percentage of correct fitting for SCAD, LASSO and AdLASSO decreases significantly more when d increases, while the performance of the TLP is relatively stable as d increases. For example, when $d = 400$, the correct fitting is 82.5% for TLP versus 58.5% for SCAD, 18% for LASSO, and 59.5% for AdLASSO in the linear spline. In addition, SCAD, LASSO and AdLASSO also tend to over-fit the model when d increases, for example, when $d = 400$, the over-fitting rate is 37% for SCAD, 81% for LASSO, and 39.5% for AdLASSO versus 14.5% for TLP in the linear spline.

In terms of estimation accuracy, Table 1 shows that the RTAISE of the TLP is close to 1 when d is small. This indicates that the TLP can estimate the nonzero components as accurately as the oracle. But RTAISE increases as d increases, since variable selection becomes more challenging as d increases. Figure 1 plots the typical estimated coefficient functions from ORACLE, TLP and SCAD using linear splines ($p = 1$) when $d = 100$. The typical estimated coefficient functions are

Penalty	d	RTAISE	C	U	O
TLP	10	1.049	0.925	0.005	0.070
SCAD		1.051	0.875	0.010	0.125
LASSO		1.080	0.640	0.000	0.360
AdLASSO		1.061	0.895	0.000	0.105
TLP	100	1.230	0.890	0.030	0.080
SCAD		1.282	0.710	0.030	0.260
LASSO		1.391	0.410	0.000	0.590
AdLASSO		1.283	0.720	0.000	0.280
TLP	200	1.404	0.895	0.035	0.070
SCAD		1.546	0.705	0.035	0.260
LASSO		1.856	0.330	0.015	0.655
AdLASSO		1.509	0.710	0.015	0.275
TLP	400	1.715	0.825	0.030	0.145
SCAD		1.826	0.585	0.045	0.370
LASSO		2.364	0.180	0.010	0.810
AdLASSO		1.879	0.595	0.010	0.395

Table 1: Simulation results for model selection based on various penalty functions: Relative total averaged integrated squared errors (RTAISEs) and the percentages of correct-fitting (C), under-fitting (U) and over-fitting (O) over 200 replications.

those with TAISE being the median of the 200 TAISEs from the simulations. Also plotted are the point-wise 95% confidence intervals from the ORACLE estimation, with the point-wise lower and upper bounds being the 2.5% and 97.5% sample quantiles of the 200 ORACLE estimates. Figure 1 shows that the proposed TLP method estimates the coefficient functions reasonably well. Compared with the SCAD, LASSO and AdLASSO, the TLP method gives better estimation in general, which is consistent with the RTAISEs reported in Table 1.

5.2 Application to AIDS Data

In this subsection, we consider the AIDs data in Huang, Wu and Zhou (2004). The data set consists of 283 homosexual males who were HIV positive between 1984 and 1991. Each patient was scheduled to undergo measurements related to their disease at a semi-annual base visit, but some of them missed or rescheduled their appointments. Therefore, each patient had different measurement times during the study period. It is known that HIV destroys CD4 cells, so by measuring CD4 cell counts and percentages in the blood, patients can be regularly monitored for disease progression. One of the study goals is to evaluate the effects of cigarette smoking status (Smoking), with 1 as smoker and 0 as nonsmoker; pre-HIV infection CD4 cell percentage (Precd4); and age at HIV infection (age), on the CD4 percentage after infection. Let t_{ij} be the time in years of the jth measurement for the ith individual after HIV infection, and y_{ij} be the CD4 percentage of patient i at time t_{ij}. We consider the following varying-coefficient model

$$y_{ij} = \beta_0(t_{ij}) + \beta_1(t_{ij})\text{Smoking} + \beta_2(t_{ij})\text{Age} + \beta_3(t_{ij})\text{Precd4} + \varepsilon_{ij}. \tag{11}$$

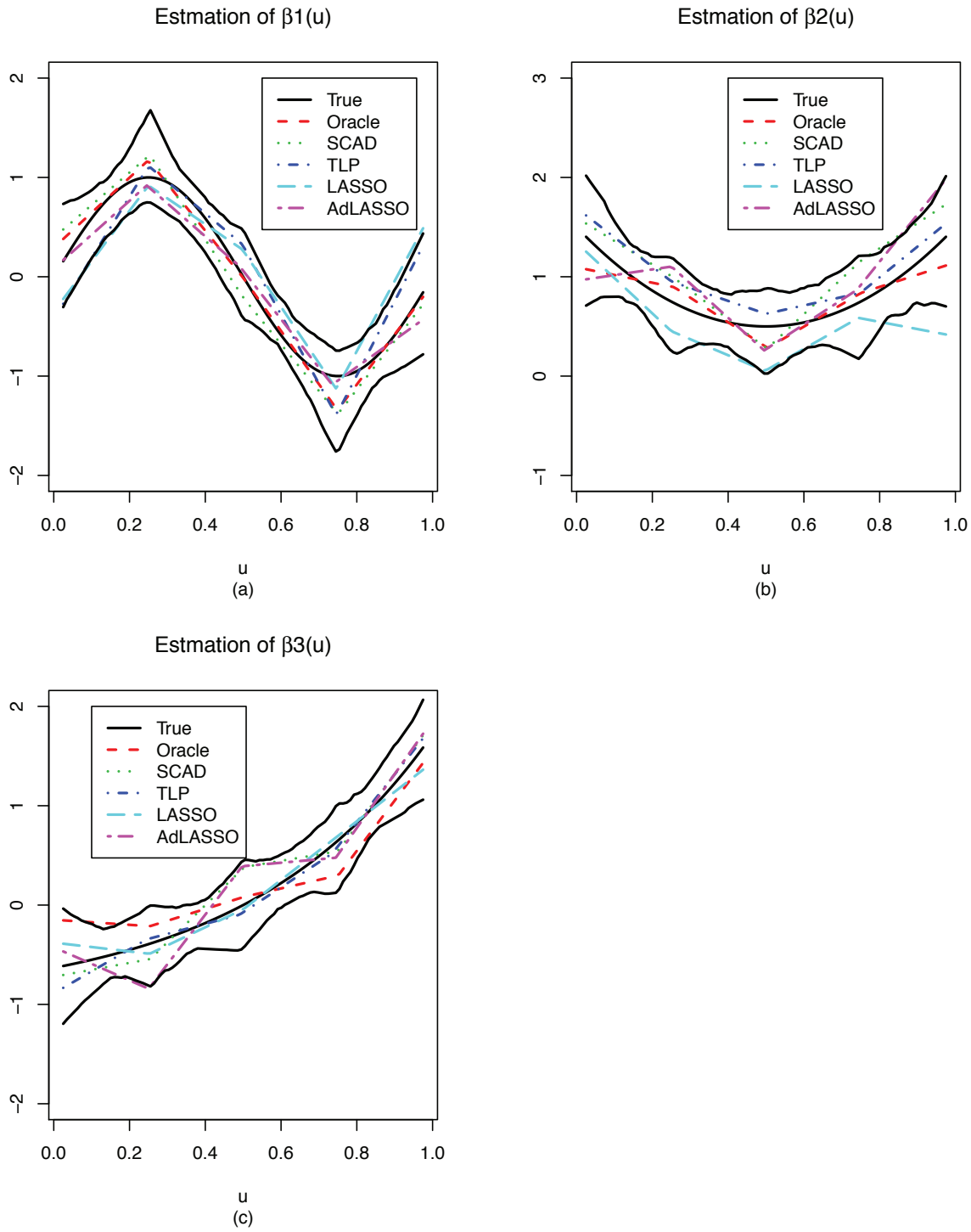

Figure 1: Simulated example: Plots of the estimated coefficient functions for (a) $\beta_1(u)$, (b) $\beta_2(u)$ and (c) $\beta_3(u)$ based on Oracle, SCAD, TLP, LASSO and AdLASSO approaches using linear spline when $d = 100$. In each plot, also plotted are the true curve and the pointwise 95% confidence intervals from the ORACLE estimation.

We apply the proposed penalized cubic spline ($p = 3$) with TLP, SCAD, LASSO and Adaptive LASSO penalties to identify the non-zero coefficient functions. We also consider the standard polynomial spline estimation of the coefficient functions. All four procedures selected two non-zero coefficient functions $\beta_0(t)$ and $\beta_3(t)$, indicating that Smoking and Age have no effect on the CD4 percentage. Figure 2 plots the estimated coefficient functions from the standard cubic spline, SCAD, TLP, LASSO and Adaptive LASSO approaches. For the standard cubic spline estimation, we also calculated the 95% point-wise bootstrap confidence intervals for the coefficient functions based on 500 bootstrapped samples.

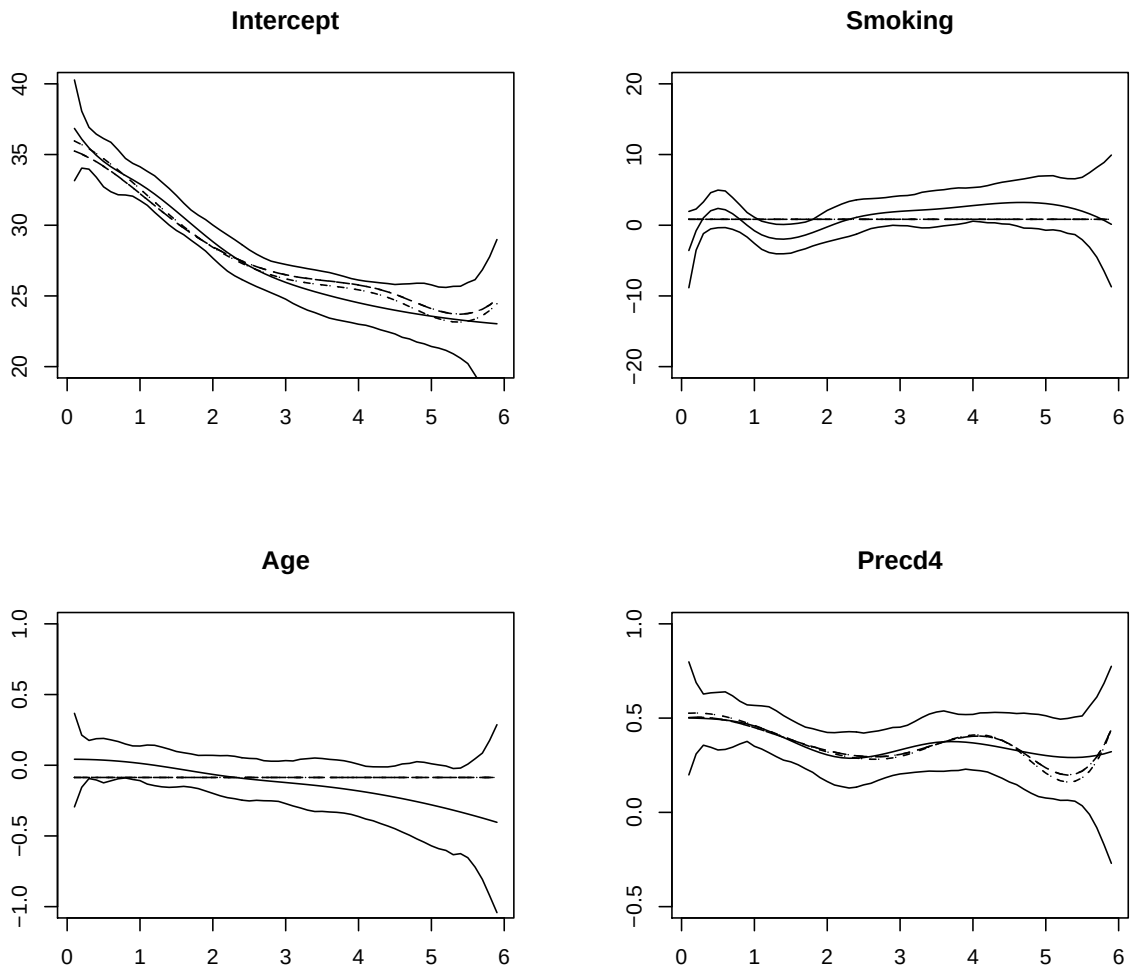

Figure 2: AIDs data: Plots of the estimated coefficient functions using standard cubic spline (line), penalized cubic spline with TLP (dotted), SCAD (dashed), LASSO (dotdash), Adaptive LASSO (long dash) penalties, together with the point-wise 95% bootstrap confidence intervals from the standard cubic spline estimation.

In this example, the dimension of the linear covariates is rather small. In order to evaluate a more challenging situation with higher dimension of d, we introduced an additional 100 redundant linear covariates, which are artificially generated from a Uniform $[0,1]$ distribution independently. We then apply the penalized spline with TLP, SCAD, LASSO or Adaptive LASSO penalties to the augmented data set. We repeated this procedure 100 times. For the three observed variables in model (11), all four procedures always select the Precd4 and never select Smoking and Age. For the 100 artificial covariates, the TLP selects at least one of these artificial covariates only 8 times, while LASSO, Adaptive LASSO, and SCAD select 28, 27, and 42 times respectively. Clearly, LASSO, Adaptive LASSO and SCAD tend to overfit the model and select many more null variables in this data example. Note that our analysis does not incorporate the dependent structure of the repeated measurements. Using the dependent structure of correlated data for high-dimensional settings will be further investigated in our future research.

6. Discussion

We propose simultaneous model selection and parameter estimation for the varying-coefficient model in high-dimensional settings where the dimension of predictors exceeds the sample size. The proposed model selection approach approximates the L_0 penalty effectively, while overcoming the computational difficulty of the L_0 penalty. The key idea is to decompose the non-convex penalty function by taking the difference between two convex functions, therefore transforming a non-convex problem into a convex optimization problem. The main advantage is that the minimization process does not depend on the initial consistent estimators of coefficients, which could be hard to obtain when the dimension of covariates is high. Our simulation and data examples confirm that the proposed model selection performs better than the SCAD in the high-dimensional case.

The model selection consistency property is derived for the proposed method. In addition, we show that it possesses the oracle property when the dimension of covariates exceeds the sample size. Note that the theoretical derivation of asymptotic properties and global optimality results are rather challenging for varying-coefficient model selection, as the dimension of the nonparametric component is also infinite in addition to the high-dimensional covariates.

Shen, Pan and Zhu (2012) provide stronger conditions under which a local minimizer can also achieve the objective of a global minimizer through the penalized truncated L_1 approach. The derivation is based on the normality assumption and the projection theory. For the nonparametric varying-coefficient model, these assumptions are not necessarily satisfied and the projection property cannot be used due to the curse of dimensionality. In general, whether a local minimizer can also hold the global optimality property for the high-dimensional varying-coefficient model requires further investigation. Nevertheless, the DC algorithm yields a better local minimizer compared to the SCAD, and can achieve the global minimum if it is combined with the branch-and-bound method (Liu, Shen and Wong, 2005), although this might be more computationally intensive.

Acknowledgments

Xue's research was supported by the National Science Foundation (DMS-0906739). Qu's research was supported by the National Science Foundation (DMS-0906660). The authors are grateful to

Xinxin Shu's computing support, and the three reviewers and the Action Editor for their insightful comments and suggestions which have improved the manuscript significantly.

Appendix A. Assumptions

To establish the asymptotic properties of the spline TLP estimators, we introduce the following notation and technical assumptions. For a given sample size n, let $\mathbf{Y}_n = (Y_1,\ldots,Y_n)^T$, $\mathbb{X}_n = (\mathbf{X}_1,\ldots,\mathbf{X}_n)^T$ and $\mathbf{U}_n = (U_1\ldots,U_n)^T$. Let $\mathbf{X}_{nj}$ be the j-th column of $\mathbb{X}_n$. Let $\|\cdot\|_2$ be the usual L_2 norm for functions and vectors and $\mathcal{C}^p([a,b])$ be the space of p-times continuously differentiable functions defined on $[a,b]$. For two vectors of the same length $a = (a_1,\ldots,a_d)^T$ and $b = (b_1,\ldots,b_d)^T$, denote $a \circ b = (a_1b_1,\ldots,a_db_d)^T$. For any scalar function $g(\cdot)$ and a vector $a = (a_1,\ldots,a_d)^T$, we denote $g(a) = (g(a_1),\ldots,g(a_d))^T$.

(C1) *The number of relevant linear covariates d_0 is fixed and there exists $\beta_{0j}(\cdot) \in C^p[a,b]$ for some $p \geq 1$ and $j = 1,\ldots,d_0$, such that $E(Y|\mathbf{X},U) = \sum_{j=1}^{d_0} \beta_{0j}(U) X_j$. Furthermore there exists a constant $c_1 > 0$ such that $\min_{1\leq j\leq d_0} E\left[\beta_{0j}^2(U)\right] > c_1$.*

(C2) *The noise ε satisfies $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2 < \infty$, and its tail probability satisfies $P(|\varepsilon| > x) \leq c_2 \exp\left(-c_3 x^2\right)$ for all $x \geq 0$ and for some positive constants c_2 and c_3.*

(C3) *The index variable U has a compact support on $[a,b]$ and its density is bounded away from 0 and infinity.*

(C4) *The eigenvalues of matrix $E\left(\mathbf{X}\mathbf{X}^T|U=u\right)$ are bounded away from 0 and infinity uniformly for all $u \in [a,b]$.*

(C5) *There exists a constant $c > 0$ such that $\left|X_j\right| < c$ with probability 1 for $j = 1,\ldots,d$.*

(C6) *The d sets of knots denoted as $\upsilon_j = \left\{a = \upsilon_{j,0} < \upsilon_{j,1} < \cdots < \upsilon_{j,N_n} < \upsilon_{j,N_n+1} = b\right\}, j = 1,\ldots,d$, are quasi-uniform, that is, there exists $c_4 > 0$, such that*

$$\max_{j=1,\ldots,d} \frac{\max\left(\upsilon_{j,l+1} - \upsilon_{j,l}, l = 0,\ldots,N_n\right)}{\min(\upsilon_{j,l+1} - \upsilon_{j,l}, l = 0,\ldots,N_n)} \leq c_4.$$

(C7) *The tuning parameters satisfy*

$$\begin{aligned} \frac{\tau_n}{\lambda_n}\sqrt{\frac{\log(N_nd)}{nN_n}} + \frac{\tau_n N_n^{-(p+2)}}{\lambda_n} &= o(1) \\ \frac{N_n \log(N_nd)}{n} + \tau_n &= o(1). \end{aligned}$$

(C8) *The tuning parameters satisfy*

$$\begin{aligned} \frac{\log(N_nd)N_n}{n\lambda_n} + \frac{n}{\log(N_nd)N_n^{2p+3}} &= o(1) \\ \frac{n\lambda_n}{\log(N_nd)dN_n} + \frac{d\log(n)\tau_n^2}{\lambda_n} &= o(1). \end{aligned}$$

(C9) *For any subset A of* $\{1,\ldots,d\}$, *let*

$$\Delta_n(A) = \min_{\beta_j \in \varphi_j, j \in A} \left\| \sum_{j \in A} \beta_j(\mathbf{U}_n) \circ \mathbf{X}_{nj} - \sum_{j \in A_0} \beta_{0j}(\mathbf{U}_n) \circ \mathbf{X}_{nj} \right\|_2^2.$$

We assume that the model (1) is empirically identifiable in the sense that,

$$\lim_{n\to\infty} \min \left\{ (\log(N_n d) N_n d)^{-1} \Delta_n(A) : A \neq A_0, |A| \leq \alpha d_0 \right\} = \infty,$$

where $\alpha > 1$ is a constant, $|A|$ *denotes the cardinality of* A, and $A_0 = \{1,\ldots,d_0\}$.

The above conditions are commonly assumed in the polynomial spline and variable selection literature. Conditions similar to (C1) and (C2) are also assumed in Huang, Horowitz and Wei (2010). Conditions similar to (C3)-(C6) can be found in Huang, Wu and Zhou (2002) and are needed for estimation consistency even when the dimension of linear covariates d is fixed. Conditions (C7) and (C8) are two different sets of conditions on tuning parameters for the local and global optimality of the spline TLP, respectively. Condition (C9) is analogous to the "degree-of-separation" condition assumed in Shen, Pan and Zhu (2012), and is weaker than the sparse Riesz condition assumed in Wei, Huang and Li (2011).

Appendix B. Outline of Proofs

To establish the asymptotic properties of the proposed estimator, we first investigate the properties of spline functions for high-dimensional data in Lemmas 4-5 and properties of the oracle spline estimators of the coefficient functions in Lemma oracle. The approximation theory for spline functions (De Boor, 2001) plays a key role in these proofs. When the true model is assumed to be known, it reduces to the estimation of the the varying-coefficient model with fixed dimensions. The asymptotic properties of the resulting oracle spline estimators of the coefficient functions have been discussed in the literature.Specifically, Lemma 6 follows directly from Theorems 2 and 3 of Huang, Wu and Zhou (2004).

To prove Theorem 1, we first provide the sufficient conditions for a solution to be a local minimizer for the object function by differentiating the objective function through regular subdifferentials. We then establish Theorem 1 by showing that the oracle estimator satisfies those conditions with probability approaching 1. In Theorem 2, we show that the oracle estimator minimizes the objective function globally with probability approaching 1, thereby establishing that the oracle estimator is also the global optimizer. This is accomplished by showing that the sum of the probabilities of all the other misspecified solutions minimizing the objective function converges to zero as $n \to \infty$.

Appendix C. Technical Lemmas

For any set $A \subset \{1,\ldots,d\}$, we denote $\widetilde{\beta}^{(A)}$ the standard polynomial spline estimator of the model A, that is, $\widetilde{\beta}_j^{(A)} = 0$ if $j \notin A$, and

$$\left(\widetilde{\beta}_j^{(A)}, j \in A\right) = \operatorname*{argmin}_{s_j \in \varphi_j} \frac{1}{2n} \sum_{i=1}^{n} \left[Y_i - \sum_{j \in A} s_j(U_i) X_{ij} \right]^2. \tag{12}$$

In particular, $\widetilde{\beta}^{(o)} = \widetilde{\beta}^{(A_0)}$, with $A_0 = \{1,\ldots,d_0\}$ being the standard polynomial spline estimator of the oracle model.

We first investigate the property of splines. Here we use B-spline basis in the proof, but the results still hold true for other choices of basis. For any $s^{(1)}(u) = \left(s_1^{(1)}(u),\ldots,s_d^{(1)}(u)\right)^T$ and $s^{(2)}(u) = \left(s_1^{(2)}(u),\ldots,s_d^{(2)}(u)\right)^T$ with each $s_j^{(1)}(u), s_j^{(2)}(u) \in S_j$, define the empirical inner product as

$$\left\langle s^{(1)}, s^{(2)} \right\rangle_n = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d} s_j^{(1)}(U_i)X_{ij}\right)\left(\sum_{j=1}^{d} s_j^{(2)}(U_i)X_{ij}\right),$$

and theoretical inner product as

$$\left\langle s^{(1)}, s^{(2)} \right\rangle = E\left[\left(\sum_{j=1}^{d} s_j^{(1)}(U)X_j\right)\left(\sum_{j=1}^{d} s_j^{(2)}(U)X_j\right)\right].$$

Denote the induced empirical and theoretical norms as $\|\cdot\|_n$ and $\|\cdot\|$ respectively. Let $\|g\|_\infty = \sup_{x\in[a,b]} g(u)$ be the supremum norm.

Lemma 4 *For any $s_j(u) \in \varphi_j$, write $s_j(u) = \sum_{l=1}^{J_n}\gamma_{jl}B_{jl}(u)$ for $\gamma_j = (\gamma_{j1},\ldots,\gamma_{jJ_n})^T$. Let $\gamma = \left(\gamma_1^T,\ldots,\gamma_d^T\right)^T$ and $\mathbf{s}(u) = (s_1(u),\ldots,s_d(u))^T$. Then there exist constants $0 < c \le C$ such that*

$$c\|\gamma\|_2^2/N_n \le \|\mathbf{s}\|^2 \le C\|\gamma\|_2^2/N_n.$$

Proof: Note that

$$\begin{aligned}\|\mathbf{s}\|^2 &= E\left[\left(\sum_{j=1}^{d} s_j(U)X_j\right)^2\right] = E\left[\mathbf{s}^T(U)\mathbf{X}\mathbf{X}^T\mathbf{s}(U)\right] \\ &= E\left[\mathbf{s}^T(U)E\left\{\mathbf{X}\mathbf{X}^T|U\right\}\mathbf{s}(U)\right].\end{aligned}$$

Therefore by (C4), there exist $0 < c_1 \le c_2$, such that

$$c_1 E\left[\mathbf{s}^T(U)\mathbf{s}(U)\right] \le \|\mathbf{s}\|^2 \le c_2 E\left[\mathbf{s}^T(U)\mathbf{s}(U)\right],$$

in which, by properties of B-spline basis functions, there exist $0 < c_1^* \le c_2^*$, such that

$$c_1^*\sum_{j=1}^{d}\left\|\gamma_j\right\|_2^2/N_n \le E\left[\mathbf{s}^T(U)\mathbf{s}(U)\right] = \sum_{j=1}^{d} E\left[s_j^2(U)\right] \le c_2^*\sum_{j=1}^{d}\left\|\gamma_j\right\|_2^2/N_n.$$

The conclusion follows by taking $c = c_1c_1^*$, and $C = c_2c_2^*$.

For any $A \subset \{1,\ldots,d\}$, let $|A|$ be the cardinality of A. Denote $\mathbf{Z}_A = (\mathbf{Z}_j, j \in A)$ and $\mathbf{D}_A = \mathbf{Z}_A^T\mathbf{Z}_A/n$. Let $\rho_{\min}(\mathbf{D}_A)$ and $\rho_{\max}(\mathbf{D}_A)$ be the minimum and maximum eigenvalues of $\mathbf{D}_A$ respectively. ■

Lemma 5 *Suppose that $|A|$ is bounded by a fixed constant independent of n and d. Then under conditions (C3)-(C5), one has*

$$c_1/N_n \le \rho_{\min}(\mathbf{D}_A) \le \rho_{\max}(\mathbf{D}_A) \le c_2/N_n,$$

for some constants $c_1, c_2 > 0$.

Proof: Without loss of generality, we assume $A = \{1, \ldots, k\}$ for some constant k which does not depend on n nor d. Note that for any $\gamma_A = (\gamma_j, j \in A)$, the triangular inequality gives

$$\gamma_A^T \mathbf{D}_A \gamma_A = \frac{1}{n} \left\| \sum_{j \in A} \mathbf{Z}_j \gamma_j \right\|_2^2 \leq \frac{2}{n} \sum_{j \in A} \left\| \mathbf{Z}_j \gamma_j \right\|_2^2 = 2 \sum_{j \in A} \gamma_j^T \mathbf{D}_j \gamma_j,$$

where $\mathbf{D}_j = \mathbf{Z}_j^T \mathbf{Z}_j / n$. By Lemma 6.2 of Zhou, Shen and Wolfe (1998), there exist constants $c_3, c_4 > 0$ $c_3/N_n \leq \rho_{\min}(\mathbf{D}_j) \leq \rho_{\max}(\mathbf{D}_j) \leq c_4/N_n$. Therefore $\gamma_A^T \mathbf{D}_A \gamma_A \leq 2c_4 \gamma_A^T \gamma_A / N_n$. That is $\rho_{\max}(\mathbf{D}_A) \leq 2c_4/N_n = c_2/N_n$. The lower bound follows from Lemma A.5 in Xue and Yang (2006) with $d_2 = 1$.

Now we consider properties of the oracle spline estimators of the coefficient functions when the true model is known. That is, $\widehat{\beta}^{(o)} = \left(\widehat{\beta}_1^{(o)}, \ldots, \widehat{\beta}_{d_0}^{(o)}, 0, \ldots, 0 \right)$ is the polynomial spline estimator of coefficient functions knowing only that the first d_0 covariates are relevant. That is

$$\left(\widehat{\beta}_1^{(o)}, \ldots, \widehat{\beta}_{d_0}^{(o)} \right)^T = \underset{s_j \in \varphi_j}{\operatorname{argmin}} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{d_0} s_j(U_i) X_{ij} \right]^2.$$

■

Lemma 6 *Suppose conditions (C1)-(C6) hold. If* $\lim N_n \log N_n / n = 0$*, then for* $j = 1, \ldots, d_0$,

$$\begin{aligned} E\left(\beta_j(U) - \widehat{\beta}_j^{(o)}(U) \right)^2 &= O_p\left(\frac{N_n}{n} + N_n^{-2(p+1)} \right), \\ \frac{1}{n} \sum_{i=1}^{n} \left(\beta_j(U_i) - \widehat{\beta}_j^{(o)}(U_i) \right)^2 &= O_p\left(\frac{N_n}{n} + N_n^{-2(p+1)} \right), \end{aligned}$$

and

$$\left\{ V\left(\widehat{\beta}^{(o,1)}(u) \right) \right\}^{-1/2} \left(\widehat{\beta}^{(o,1)}(u) - \beta^{(1)}(u) \right) \to N(0, \mathbf{I})$$

in distribution, where $\widehat{\beta}^{(o,1)}(u) = \left(\widehat{\beta}_1^{(o)}(u), \ldots, \widehat{\beta}_{d_0}^{(o)}(u) \right)^T$*, and* $\beta^{(1)}(u) = (\beta_1(u), \ldots, \beta_{d_0}(u))^T$*, and*

$$V\left(\widehat{\beta}^{(o,1)}(u) \right) = \mathbf{B}^{(1)}(u) \left(\sum_{i=1}^{n} \mathbf{A}_i^{(1)T} \mathbf{A}_i^{(1)} \right)^{-1} \mathbf{B}^{(1)}(u) = O_p(N_n/n),$$

where $\mathbf{B}^{(1)}(u) = \left(\mathbf{B}_1^T(u), \ldots, \mathbf{B}_{d_0}^T(u) \right)^T$*, and* $\mathbf{A}_i^{(1)} = \left(\mathbf{B}_1^T(U_i) X_{i1}, \ldots, \mathbf{B}_{d_0}^T(U_i) X_{id_0} \right)^T$ *in which* $\mathbf{B}_j^T(U_i) X_{ij} = (B_{j1}(U_i) X_{ij}, \ldots, B_{jJ_n}(U_i) X_{ij})$.

Proof: It follows from Theorems 2 and 3 of Huang, Wu and Zhou (2004). ■

Lemma 7 *Suppose conditions (C1)-(C6) hold. Let* $T_{jl} = \sqrt{N_n/n} \sum_{i=1}^{n} B_{jl}(U_i) X_{ij} \varepsilon_i$*, for* $j = 1, \ldots, d$*, and* $l = 1, \ldots, J_n$*. Let* $T_n = \max_{1 \leq j \leq d, 1 \leq l \leq J_n} |T_{jl}|$*. If* $N_n \log(N_n d)/n \to 0$*, then*

$$E(T_n) = O\left(\sqrt{\log(N_n d)} \right).$$

Proof: Let $m_{jl}^2 = \sum_{i=1}^{n} B_{jl}^2(U_i) X_{ij}^2$, and $m_n^2 = \max_{1\le j\le d, 1\le l\le J_n} m_{jl}^2$. By condition (C2) and the maximal inequality for gaussian random variables, there exists a constant $C_1 > 0$ such that

$$E(T_n) = E\left(\max_{1\le j\le d, 1\le l\le J_n} |T_{jl}|\right) \le C_1\sqrt{N_n/n}\sqrt{\log(N_n d)}E(m_n). \tag{13}$$

Furthermore, by the definition of B-spline basis and (C5), there exists a $C_2 > 0$, such that for each $1 \le j \le d, 1 \le l \le J_n$,

$$\left|B_{jl}^2(U_i) X_{ij}^2\right| \le C_2, \text{ and } E\left[B_{jl}^2(U_i) X_{ij}^2\right] \le C_2 N_n^{-1}.$$

As a result,

$$\sum_{i=1}^{n} E\left[B_{jl}^2(U_i) X_{ij}^2 - E\left(B_{jl}^2(U_i) X_{ij}^2\right)\right]^2 \le 4C_2 n N_n^{-1},$$

and

$$\max_{1\le j\le d, 1\le l\le J_n} Em_{jl}^2 = \max_{1\le j\le d, 1\le l\le J_n} \sum_{i=1}^{n} E\left(B_{jl}^2(U_i) X_{ij}^2\right) \le C_2 n N_n^{-1}. \tag{14}$$

Then by Lemma A.1 of Van de Geer (2008), one has

$$\begin{aligned} & E\left(\max_{1\le j\le d, 1\le l\le J_n} \left|m_{jl}^2 - Em_{jl}^2\right|\right) \\ = \; & E\left(\max_{1\le j\le d, 1\le l\le J_n} \left|\sum_{i=1}^{n} B_{jl}^2(U_i) X_{ij}^2 - E\left(B_{jl}^2(U_i) X_{ij}^2\right)\right|\right) \\ \le \; & \sqrt{2C_2 n N_n^{-1}\log(N_n d)} + 4\log(2N_n d). \end{aligned} \tag{15}$$

Therefore (14) and (15) give that

$$\begin{aligned} Em_n^2 &\le \max_{1\le j\le d, 1\le l\le J_n} Em_{jl}^2 + E\left(\max_{1\le j\le d, 1\le l\le J_n} \left|m_{jl}^2 - Em_{jl}^2\right|\right) \\ &\le C_2 n N_n^{-1} + \sqrt{2C_2 n N_n^{-1}\log(N_n d)} + 4\log(2N_n d). \end{aligned}$$

Furthermore, $Em_n \le \sqrt{Em_n^2} \le \left(\sqrt{2C_2 n N_n^{-1}\log(N_n d)} + 4\log(2dN_n) + C_2 n N_n^{-1}\right)^{1/2}$. Together with (13) and $N_n \log(N_n d)/n \to 0$, one has

$$\begin{aligned} E(T_n) &\le C_1\sqrt{N_n/n}\sqrt{\log(N_n d)}\left(\sqrt{2C_2 n N_n^{-1}\log(N_n d)} + 4\log(2N_n d) + C_2 n N_n^{-1}\right)^{1/2} \\ &= O\left(\sqrt{\log(N_n d)}\right). \end{aligned}$$

■

Lemma 8 *Suppose conditions (C1)-(C7) hold. Let* $\mathbf{Z}_j = (\mathbf{Z}_{1j}, \ldots, \mathbf{Z}_{nj})^T$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, *and* $\mathbf{Z}_{(1)} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_{d_0})$. *Then*

$$P\left(\left\|\frac{1}{n}\mathbf{Z}_j^T\left(\mathbf{Y} - \mathbf{Z}_{(1)}\widehat{\gamma}^{(o,1)}\right)\right\|_{W_j} > \frac{\lambda_n}{\tau_n}, \; \exists\, j = d_0+1, \ldots, d\right) \to 0.$$

Proof: By the approximation theory (de Boor 2001, p. 149), there exist a constant $c > 0$ and spline functions $s_j^0 = \sum_{l=1}^{J_n} \gamma_{jl}^0 B_{jl}(t) \in S_j$, such that

$$\max_{1 \le j \le d_0} \left\| \beta_j - s_j^0 \right\|_\infty \le c N_n^{-(p+1)}. \tag{16}$$

Let $\delta_i = \sum_{j=1}^{d_0} \left[\beta_j(U_i) - s_j^0(U_i) \right] X_{ij}$, $\delta = (\delta_1, \ldots, \delta_n)^T$, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Then one has

$$\mathbf{Z}_j^T \left(\mathbf{Y} - \mathbf{Z}_{(1)} \widehat{\gamma}^{(o,1)} \right) = \mathbf{Z}_j^T \mathbf{H}_n \mathbf{Y} = \mathbf{Z}_j^T \mathbf{H}_n \varepsilon + \mathbf{Z}_j^T \mathbf{H}_n \delta,$$

where $\mathbf{H}_n = \mathbf{I} - \mathbf{Z}_{(1)} \left(\mathbf{Z}_{(1)}^T \mathbf{Z}_{(1)} \right)^{-1} \mathbf{Z}_{(1)}^T$. By Lemma 7, there exists a $c > 0$ such that

$$E \left(\max_{d_0+1 \le j \le d} \left\| \mathbf{Z}_j^T \mathbf{H}_n \varepsilon \right\|_{W_j} \right) \le c \sqrt{n \log(N_n d) / N_n}.$$

Therefore by Markov's inequality, one has

$$\begin{aligned} & P \left(\left\| \mathbf{Z}_j^T \mathbf{H}_n \varepsilon \right\|_{W_j} > \frac{n \lambda_n}{2 \tau_n}, \ \exists \ j = d_0 + 1, \ldots, d \right) = P \left(\max_{d_0+1 \le j \le d} \left\| \mathbf{Z}_j^T \mathbf{H}_n \varepsilon \right\|_{W_j} > \frac{n \lambda_n}{2 \tau_n} \right) \\ \le & \quad \frac{2c\tau_n}{\lambda_n} \sqrt{\frac{\log(N_n d)}{n N_n}} \to 0, \end{aligned} \tag{17}$$

as $n \to \infty$, by condition (C7). On the other hand, let ρ_j and ρ_{H_n} be the largest eigenvalue of $\mathbf{Z}_j^T \mathbf{Z}_j / n$ and $\mathbf{H}_n$. Then Lemma (5) entails that $\max_{d_0+1 \le j \le d} \rho_j = O_p(1/N_n)$. Together with (16) and condition (C7), one has

$$\begin{aligned} \max_{d_0+1 \le j \le d} \frac{1}{n} \left\| \mathbf{Z}_j^T \mathbf{H}_n \delta \right\|_{W_j} & \le (n N_n)^{-1/2} \sqrt{\max_{d_0+1 \le j \le d} \rho_j \rho_{H_n}} \|\delta\|_2 \\ & = O_p \left(N_n^{-(p+1)} / N_n \right) = o_p \left(\frac{\lambda_n}{2\tau_n} \right). \end{aligned} \tag{18}$$

Then the lemma follows from (17) and (18) and by noting that

$$\begin{aligned} & P \left(\left\| \frac{1}{n} \mathbf{Z}_j^T \left(\mathbf{Y} - \mathbf{Z}_{(1)} \widehat{\gamma}^{(o,1)} \right) \right\|_{W_j} > \frac{\lambda_n}{\tau_n}, \ \exists \ j = d_0 + 1, \ldots, d \right) \\ \le & \quad P \left(\max_{d_0+1 \le j \le d} \frac{1}{n} \left\| \mathbf{Z}_j^T \mathbf{H}_n \varepsilon \right\|_{W_j} > \frac{\lambda_n}{2\tau_n} \right) + P \left(\max_{d_0+1 \le j \le d} \frac{1}{n} \left\| \mathbf{Z}_j^T \mathbf{H}_n \delta \right\|_{W_j} > \frac{\lambda_n}{2\tau_n} \right). \end{aligned}$$

■

Appendix D. Proof of Theorem 1

For notation simplicity, let $\mathbf{Z}_{ij}^* = \mathbf{W}_j^{-1/2} \mathbf{Z}_{ij}$ and $\gamma_j^* = \mathbf{W}_j^{1/2} \gamma_j$. Then the minimization problem in (4) becomes

$$L_n(\gamma^*) = \frac{1}{2n} \sum_{i=1}^n \left[Y_i - \sum_{j=1}^d \gamma_j^{*T} \mathbf{Z}_{ij}^* \right]^2 + \lambda_n \sum_{j=1}^d p_n \left(\left\| \gamma_j^* \right\|_2 \right).$$

For $i=1,\ldots,n$, and $j=1,\ldots,d$, write $\mathbf{Z}_i^*=\left(\mathbf{Z}_{i1}^{*T},\ldots,\mathbf{Z}_{id}^{*T}\right)^T$, $\gamma^*=\left(\gamma_1^{*T},\ldots,\gamma_d^{*T}\right)^T$ and $c_j^*(\gamma^*)=-\frac{1}{n}\sum_{i=1}^n\mathbf{Z}_{ij}^*\left(Y_i-\mathbf{Z}_i^{*T}\gamma^*\right)$. Differentiate $L_n(\gamma^*)$ with respect to γ_j^* through regular subdifferentials, we obtain the local optimality condition for $L_n(\gamma^*)$ as $c_j^*(\gamma^*)+\frac{\lambda_n}{\tau_n}\zeta_j=\mathbf{0}$, where $\zeta_j=\gamma_j^*/\left\|\gamma_j^*\right\|_2$ if $0<\left\|\gamma_j^*\right\|_2<\tau_n$; $\zeta_j=\{\gamma_j^*,\left\|\gamma_j^*\right\|_2\le 1\}$ if $\left\|\gamma_j^*\right\|_2=0$;$\zeta_j=\mathbf{0}$, if $\left\|\gamma_j^*\right\|_2>\tau_n$; and $\zeta_j=\emptyset$, if $\left\|\gamma_j^*\right\|_2=\tau_n$, where $\emptyset$ is an empty set. Therefore any γ^* that satisfies

$$\begin{aligned} c_j^*(\gamma^*) &= \mathbf{0}, \qquad \left\|\gamma_j^*\right\|>\tau_n \text{ for } j=1,\ldots,d_0.\\ \left\|c_j^*(\gamma^*)\right\|_2 &\le \frac{\lambda_n}{\tau_n}, \qquad \left\|\gamma_j^*\right\|=0 \text{ for } j=d_0+1,\ldots,d, \end{aligned}$$

is a local minimizer of $L_n(\gamma^*)$. Or equivalently, any γ that satisfies

$$c_j(\gamma) = \mathbf{0}, \qquad \left\|\gamma_j\right\|_{W_j}>\tau_n \text{ for } j=1,\ldots,d_0. \tag{19}$$

$$\left\|c_j(\gamma)\right\|_{W_j} \le \frac{\lambda_n}{\tau_n}, \qquad \left\|\gamma_j\right\|_{W_j}=0 \text{ for } j=d_0+1,\ldots,d, \tag{20}$$

is a local minimizer of $L_n(\gamma)$, in which $c_j(\gamma)=-\frac{1}{n}\sum_{i=1}^n\mathbf{Z}_{ij}\left(Y_i-\mathbf{Z}_i^T\gamma\right)$. Therefore it suffices to show that $\widehat{\gamma}^{(o)}$ satisfies (19) and (20).

For $j=1,\ldots,d_0$, $c_j\left(\widehat{\gamma}^{(o)}\right)=\mathbf{0}$ trivially by the definition of $\widehat{\gamma}^{(o)}$. On the other hand, conditions (C1), (C7) and Lemma 6 give that

$$\lim_{n\to\infty}P\left(\left\|\widehat{\gamma}_j^{(o)}\right\|_{W_j}>\tau_n, j=1,\ldots,d_0\right)=1.$$

Therefore $\widehat{\gamma}^{(o)}$ satisfies (19). For (20), note that, by definition $\widehat{\gamma}_j^{(o)}=0$, for $j=d_0+1,\ldots,d$. Furthermore, for $j=d_0+1,\ldots,d$,

$$c_j\left(\widehat{\gamma}^{(o)}\right)=-\frac{1}{n}\mathbf{Z}_j^T\left(\mathbf{Y}-\mathbf{Z}_{(1)}\widehat{\gamma}^{(o,1)}\right).$$

By Lemma 8,

$$P\left(\left\|c_j\left(\widehat{\gamma}^{(o)}\right)\right\|_{W_j}>\frac{\lambda_n}{\tau_n},\ \exists\ j=d_0+1,\ldots,d\right)\to 0.$$

Therefore $\widehat{\gamma}_j^{(o)}$ also satisfies (20) with probability approaching to 1. As a result, $\widehat{\gamma}^{(o)}$ is a local minimum of $L_n(\gamma)$ with probability approaching to 1. ■

Appendix E. Proof of Theorem 2

Note that for any $\gamma=\left(\gamma_1^T,\ldots,\gamma_d^T\right)^T$, one can write

$$\begin{aligned} L_n(\gamma) &= \frac{1}{2n}\sum_{i=1}^n\left[Y_i-\sum_{j=1}^d\gamma_j^T\mathbf{Z}_{ij}\right]^2+\lambda_n\sum_{j=1}^d\min\left(\left\|\gamma_j\right\|_{W_j}/\tau_n,1\right)\\ &= \frac{1}{2n}\sum_{i=1}^n\left[Y_i-\sum_{j=1}^d\gamma_j^T\mathbf{Z}_{ij}\right]^2+\lambda_n|A|+\frac{\lambda_n}{\tau_n}\sum_{j\in A^c}\left\|\gamma_j\right\|_{W_j}, \end{aligned}$$

where $A = A(\gamma) = \left\{ j : \left\|\gamma_j\right\|_{W_j} \geq \tau_n \right\}$, $A^c = \left\{ j : \left\|\gamma_j\right\|_{W_j} < \tau_n \right\}$, and $|A|$ denotes the cardinality of A. For a given set A, let $\widetilde{\gamma}^{(A)}$ be the coefficient from the standard polynomial spline estimation of the model A as defined in (12). Then for $a = \lambda_n / \left(d\tau_n^2 \log n\right) + 1 > 1$, one has

$$\begin{aligned}
& L_n(\gamma) - \lambda_n |A| \\
= & \frac{1}{2n} \sum_{i=1}^{n} \left[Y_i - \sum_{j \in A} \gamma_j^T \mathbf{Z}_{ij} - \sum_{j \in A^c} \gamma_j^T \mathbf{Z}_{ij} \right]^2 + \frac{\lambda_n}{\tau_n} \sum_{j \in A^c} \left\|\gamma_j\right\|_{W_j} \\
\geq & \frac{a-1}{2an} \sum_{i=1}^{n} \left[Y_i - \sum_{j \in A} \gamma_j^T \mathbf{Z}_{ij} \right]^2 - \frac{a-1}{2n} \sum_{i=1}^{n} \left[\sum_{j \in A^c} \gamma_j^T \mathbf{Z}_{ij} \right]^2 + \frac{\lambda_n}{\tau_n} \sum_{j \in A^c} \left\|\gamma_j\right\|_{W_j} \\
\geq & \frac{a-1}{2an} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{d} \widetilde{\gamma}_j^{(A)T} \mathbf{Z}_{ij} \right]^2 - \frac{d(a-1)}{2n} \sum_{i=1}^{n} \sum_{j \in A^c} \left(\gamma_j^T \mathbf{Z}_{ij}\right)^2 + \frac{\lambda_n}{\tau_n} \sum_{j \in A^c} \left\|\gamma_j\right\|_{W_j} \\
\geq & \frac{a-1}{2an} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{d} \widetilde{\gamma}_j^{(A)T} \mathbf{Z}_{ij} \right]^2 + \left(\frac{\lambda_n}{\tau_n} - \frac{a-1}{2} d\tau_n \right) \sum_{j \in A^c} \left\|\gamma_j\right\|_{W_j}.
\end{aligned}$$

Note that $\frac{\lambda_n}{\tau_n} - \frac{a-1}{2} d\tau_n > 0$ for sufficiently large n by the definition of a. Therefore,

$$L_n(\gamma) \geq \frac{a-1}{2an} \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{d} \widetilde{\gamma}_j^{(A)T} \mathbf{Z}_{ij} \right]^2 + \lambda_n |A|. \tag{21}$$

Let $\Gamma_1 = \{A : A \subset \{1,\ldots,d\}, A_0 \subset A, \text{ and } A \neq A_0\}$ be the set of overfitting models and $\Gamma_2 = \{A : A \subset \{1,\ldots,d\}, A_0 \not\subset A \text{ and } A \neq A_0\}$ be the set of underfitting models. For any γ, $A(\gamma)$ must fall into one of Γ_j, $j = 1,2$. We now show that

$$\sum_{A \in \Gamma_j} P\left(\min_{\gamma : \mathbf{A}(\gamma) = A} L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) \leq 0 \right) \to 0,$$

as $n \to \infty$, for $j = 1,2$.

Let $\mathbf{Z}(A) = (\mathbf{Z}_j, j \in A)$ and $H_n(A) = \mathbf{Z}(A)\left[\mathbf{Z}^T(A)\mathbf{Z}(A)\right]^{-1}\mathbf{Z}(A)$. Let $\mathbf{E} = (\varepsilon_1,\ldots,\varepsilon_n)^T$, $\mathbf{Y} = (Y_1,\ldots,Y_n)^T$, $m(\mathbf{X}_i, U_i) = \sum_{j=1}^{d} \beta_j(U_i) X_{ij}$ and $\mathbf{M} = (m(\mathbf{X}_1,U_1),\ldots,m(\mathbf{X}_n,U_n))^T$. Lemma 6 entails that $P\left(\min_{j=1,\ldots,d_0} \left\|\widetilde{\gamma}_j^{(o)}\right\|_{W_j} \geq \tau_n \right) \to 1$, as $n \to \infty$. Therefore it follows from (21) that, with probability approaching to one,

$$\begin{aligned}
& 2n\left\{ L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) - \lambda_n(|A| - d_0) \right\} \\
\geq & -\mathbf{Y}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{Y} - \frac{1}{a}\mathbf{Y}^T\left(\mathbf{I}_n - \mathbf{H}_n(A)\right)\mathbf{Y} \\
= & -\mathbf{E}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{E} - \mathbf{M}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{M} \\
& -2\mathbf{E}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{M} - \frac{1}{a}\mathbf{Y}^T\left(\mathbf{I}_n - \mathbf{H}_n(A)\right)\mathbf{Y} \\
= & -\mathbf{E}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{E} + I_{n1} + I_{n2} + I_{n3}.
\end{aligned}$$

Let $r(A)$ and $r(A_0)$ be the ranks of $\mathbf{H}_n(A)$ and $\mathbf{H}_n(A_0)$ respectively, and $I_n = I_{n1} + I_{n2} + I_{n3}$. Also note that if $T_m \sim \chi^2_m$, then the Cramer-Chernoff bound gives that $P(T_m - m > km) \leq \exp\left\{-\frac{m}{2}(k - \log(1+k))\right\}$ for some constant $k > 0$. Then one has,

$$\begin{aligned}
& P\left\{L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) < 0\right\} \\
= \; & P\left\{\mathbf{E}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{E} > I_n + 2n\lambda_n\left(|A| - d_0\right)\right\} \\
= \; & P\left\{\chi^2_{r(A)-r(A_0)} > I_n + 2n\lambda_n\left(|A| - d_0\right)\right\} \\
\leq \; & \exp\left\{-\frac{r(A) - r(A_0)}{2}\left[\frac{I_n + 2n\lambda_n\left(|A| - d_0\right)}{r(A) - r(A_0)} - 1 - \log\frac{I_n + 2n\lambda_n\left(|A| - d_0\right)}{r(A) - r(A_0)}\right]\right\} \\
\leq \; & \exp\left\{-\frac{r(A) - r(A_0)}{2}\left[\frac{I_n + 2n\lambda_n\left(|A| - d_0\right)}{r(A) - r(A_0)} - 1\right]\frac{1+c}{2}\right\} \qquad (22)
\end{aligned}$$

for some $0 < c < 1$. To bound (22), we consider the following two cases. Case 1 (overfitting): $A = A(\gamma) \in \Gamma_1$. Let $k = |A| - d_0$. By the spline approximation theorem (de Boor, 2001), there exist spline functions $s_j \in \varphi_j$ and constant c such that $\max_{1 \leq j \leq d_0} \left\|\beta_j - s_j\right\|_\infty \leq cN_n^{-(p+1)}$. Let $m^*(\mathbf{X}, U) = \sum_{j=1}^{d_0} s_j(U) X_j$, and $\mathbf{M}^* = \left(m^*(\mathbf{X}_1, U_1), \ldots, m^*(\mathbf{X}_n, U_n)\right)^T$. Then by the definition of projection

$$\frac{1}{n}\mathbf{M}^T\left(\mathbf{I}_n - \mathbf{H}_n(A_0)\right)\mathbf{M} \leq \|m - m^*\|_n^2 \leq cd_0N_n^{-2(p+1)}.$$

Similarly, one can show $\frac{1}{n}\mathbf{M}^T\left(\mathbf{I}_n - \mathbf{H}_n(A)\right)\mathbf{M} \leq c|A|N_n^{-2(p+1)}$. Therefore, by condition (C8)

$$I_{n1} = \mathbf{M}^T\left(\mathbf{I}_n - \mathbf{H}_n(A)\right)\mathbf{M} - \mathbf{M}^T\left(\mathbf{I}_n - \mathbf{H}_n(A_0)\right)\mathbf{M} \leq ckN_n^{-2(p+1)}n = o_p\left(k\log\left(dN_n\right)N_n\right).$$

Furthermore, the Cauchy-Schwartz inequality gives that,

$$\begin{aligned}
|I_{n2}| \; & \leq \; 2\sqrt{\mathbf{E}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{E}}\sqrt{\mathbf{M}^T\left(\mathbf{H}_n(A) - \mathbf{H}_n(A_0)\right)\mathbf{M}} \\
& = \; O_p\left(k\sqrt{\log\left(dN_n\right)N_n}nN_n^{-(p+1)}\right) = o_p\left(k\log\left(dN_n\right)N_n\right).
\end{aligned}$$

Finally $I_{n3} = -\frac{1}{a}\mathbf{Y}^T\left(\mathbf{I}_n - \mathbf{H}_n(A)\right)\mathbf{Y} = o_p\left(k\log\left(dN_n\right)N_n\right)$, since $a \to \infty$ as $n \to \infty$ by condition (C8). Therefore, $I_n = I_{n1} + I_{n2} + I_{n3} = o_p\left(k\log\left(dN_n\right)N_n\right)$. As a result, (22) gives that,

$$\begin{aligned}
& \sum_{A(\gamma)\in\Gamma_1} P\left(\min_\gamma L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) \leq 0\right) \\
\leq \; & \sum_{k=1}^{d-d_0}\left(\begin{array}{c} d-d_0 \\ k \end{array}\right)\exp\left\{-\frac{r(A) - r(A_0)}{2}\left[\frac{I_n + 2n\lambda_n k}{r(A) - r(A_0)} - 1\right]\frac{1+c}{2}\right\} \\
\leq \; & \sum_{k=1}^{d-d_0} d^k \exp\left\{-\frac{1+c}{4}\left[I_n + 2n\lambda_n k - \left(r(A) - r(A_0)\right)\right]\right\} \\
= \; & \sum_{k=1}^{d-d_0} \exp\left\{-\frac{1+c}{4}\left[I_n + 2n\lambda_n k - \left(r(A) - r(A_0)\right)\right] + k\log d\right\}
\end{aligned}$$

in which $2n\lambda_n k$ is the dominated term inside of the exponential under condition (C8). Therefore,

$$\begin{aligned}
&\sum_{A(\gamma)\in\Gamma_1} P\left(\min_{\gamma} L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) \le 0\right) \\
\le\ & \sum_{k=1}^{d-d_0} \exp\left\{-\frac{n\lambda_n k}{2}\right\} = \exp\left\{-\frac{n\lambda_n}{2}\right\} \frac{1-\exp\left(-\frac{n(d-d_0)\lambda_n}{2}\right)}{1-\exp\left(-\frac{n\lambda_n}{2}\right)} \to 0
\end{aligned} \tag{23}$$

as $n\to\infty$, by condition (C8).

Case 2 (underfitting): $A = A(\gamma) \in \Gamma_2$. Note that,

$$I_{n1} = \mathbf{M}^T(\mathbf{I}_n - \mathbf{H}_n(A))\mathbf{M} - \mathbf{M}^T(\mathbf{I}_n - \mathbf{H}_n(A_0))\mathbf{M} = I_{n1}^{(1)} - I_{n1}^{(2)},$$

in which

$$I_{n1}^{(1)} = \mathbf{M}^T(\mathbf{I}_n - \mathbf{H}_n(A))\mathbf{M} \ge \Delta_n(A).$$

Therefore for any γ with $A_0 \not\subset A$ and $|A| \le \alpha d_0$ where $\alpha > 1$ is a constant as given in condition (C9), the empirically identifiable condition entails that, $(\log(N_n d) N_n d)^{-1} I_{n1}^{(1)} \to \infty$,as $n\to\infty$. On the other hand, similar arguments for Case 1 give that $I_{n1}^{(2)} = O_p\left(d_0 N_n^{-2(p+1)} n\right) = o_p(\log(N_n d) N_n d)$, and $I_{n2} + I_{n3} = O_p(\log(N_n d) N_n d)$. Therefore $I_{n1}^{(1)}$ is the dominated term in I_n. As a result, together with (22), one has

$$P\left\{L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) < 0\right\} \le \exp\left\{-\frac{1+c}{4}\left[\frac{I_{n1}^{(1)}}{2} + 2n\lambda_n(|A| - d_0) - (r(A) - r(A_0))\right]\right\}.$$

Furthermore, note that for n large enough,

$$\begin{aligned}
2n\lambda_n(|A| - d_0) - (r(A) - r(A_0)) &\ge (2n\lambda_n - N_n - p - 1)(|A| - d_0) \\
&\ge n\lambda_n(|A| - d_0) \ge -n\lambda_n d_0 = o(\log(N_n d) N_n d)
\end{aligned}$$

by assumption (C8). Therefore $I_{n1}^{(1)}$ is the dominated term inside of the exponential. Thus, when n is large enough, one has,

$$P\left\{L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) < 0\right\} \le \exp\left\{-\frac{I_{n1}^{(1)}}{8}\right\} \le \exp\left\{-\frac{\Delta_n(A)}{8}\right\}. \tag{24}$$

For any γ with $A_0 \not\subset A$ and $|A| > \alpha d_0$, we show that, $I_n = L_1(A) + L_2(A) + L_3(A)$, where $L_1(A) = -\frac{1}{a}(\mathbf{E} - (a-1)(\mathbf{I}_n - \mathbf{H}_n(A))\mathbf{M})^T(\mathbf{I}_n - \mathbf{H}_n(A))(\mathbf{E} - (a-1)(\mathbf{I}_n - \mathbf{H}_n(A))\mathbf{M})$,
$L_2(A) = (a-1)\mathbf{M}^T(\mathbf{I}_n - \mathbf{H}_n(A))\mathbf{M}$, and

$$L_3(A) = -\mathbf{M}^T(\mathbf{I}_n - \mathbf{H}_n(A_0))\mathbf{M} - 2\mathbf{E}^T(\mathbf{I}_n - \mathbf{H}_n(A_0))\mathbf{M}.$$

Here, $-aL_1(A)/\sigma^2$ follows a noncentral χ^2 distribution with the degree of freedom $n - \min(r(A), n)$ and noncentral parameter $(a-1)\mathbf{M}^T(\mathbf{I}_n - \mathbf{H}_n(A))\mathbf{M}/\sigma^2$. Furthermore, as in Case 1, one can show

that $L_3(A) = o_p(\log(dN_n)N_n d_0)$. Therefore $L_2(A)$ is the dominated term in I_n, by noting that $a \to \infty$ by assumption (C8). Thus, for n sufficiently large,

$$\begin{aligned} & P\left\{L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) < 0\right\} \\ \leq & \exp\left\{-\frac{1+c}{4}\left[I_n + 2n\lambda_n(|A| - d_0) - (r(A) - r(A_0))\right]\right\} \\ \leq & \exp\left\{-\frac{1+c}{4}\left[2n\lambda_n(|A| - d_0) - (r(A) - r(A_0))\right]\right\}. \end{aligned} \tag{25}$$

Therefore, (24) and (25) give that,

$$\begin{aligned} & \sum_{A(\gamma)\in\Gamma_2} P\left(\min_{\gamma} L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) \leq 0\right) \\ \leq & \sum_{i=1}^{[\alpha d_0]}\sum_{j=0}^{d_0-1}\left(\begin{array}{c} d_0 \\ j \end{array}\right)\left(\begin{array}{c} d-d_0 \\ i-j \end{array}\right)\exp\left\{-\frac{\min\Delta_n(A)}{8}\right\} \\ & + \sum_{i=[\alpha d_0]+1}^{d}\sum_{j=0}^{d_0-1}\left(\begin{array}{c} d_0 \\ j \end{array}\right)\left(\begin{array}{c} d-d_0 \\ i-j \end{array}\right)\exp\left\{-\frac{1+c}{4}\left[2n\lambda_n(i-d_0) - (r(A) - r(A_0))\right]\right\} \\ = & II_1 + II_2, \end{aligned}$$

where, by noting that $\binom{a}{b} \leq a^b$ for any two integers $a, b > 0$,

$$\begin{aligned} II_1 \leq & \sum_{i=1}^{[\alpha d_0]}\sum_{j=0}^{d_0-1} d_0^j (d-d_0)^{i-j}\exp\left(-\frac{\min\Delta_n(A)}{8}\right) \\ \leq & (N_n d)^{-N_n d/8} d_0^{[\alpha d_0]}(d-d_0)^{[\alpha d_0]}[\alpha d_0] d_0 \to 0, \end{aligned}$$

as $n \to \infty$, since d_0 is fixed and $N_n \to \infty$. Furthermore,

$$\begin{aligned} II_2 \leq & \sum_{i=[\alpha d_0]+1}^{d}\sum_{j=0}^{d_0-1}\left(\begin{array}{c} d_0 \\ j \end{array}\right)\left(\begin{array}{c} d-d_0 \\ i-j \end{array}\right)\exp\left\{-\frac{1+c}{4}\left[2n\lambda_n(i-d_0) - (r(A) - r(A_0))\right]\right\} \\ \leq & \sum_{i=[\alpha d_0]+1}^{d}\sum_{j=0}^{d_0-1} d_0^j (d-d_0)^{i-j}\exp\left\{-\frac{n\lambda_n(i-d_0)}{4}\right\} \\ \leq & \sum_{i=[\alpha d_0]+1}^{d} d_0\exp\left\{-\frac{n\lambda_n(i-d_0)}{4} + i\log(d)\right\} \to 0, \end{aligned}$$

as $n \to \infty$, by assumption (C8). Therefore, as $n \to \infty$,

$$\sum_{A\in\Gamma_2} P\left(\min_{\gamma:\mathbf{A}(\gamma)=A} L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) \leq 0\right) \to 0. \tag{26}$$

Note that for the global minima $\widehat{\gamma}$ of (4), one has

$$P\left(\widehat{\gamma} \neq \widetilde{\gamma}^{(o)}\right) \leq \sum_{j=1}^{2}\sum_{A\in\Gamma_j} P\left(\min_{\gamma:\mathbf{A}(\gamma)=A} L_n(\gamma) - L_n\left(\widetilde{\gamma}^{(o)}\right) \leq 0\right).$$

Therefore, Theorem 2 follows from (23) and (26). ■

Appendix F. Proof of Theorem 3

Theorem 3 follows immediately from Lemma 6 and Theorem 2. ■

References

L. An and P. Tao. Solving a class of linearly constrained indefinite quadratic problems by D.C. algorithms. *Journal of Global Optimization*, 11:253-285, 1997.

L. Breiman and A. Cutler. A deterministic algorithm for global optimization. *Mathematical Programming*, 58:179-199, 1993.

C. de Boor. *A Practical Guide to Splines*. Springer, New York, 2001.

J. Fan and T. Huang. Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11:1031-1057, 2005.

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

J. Fan and H. Peng. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics*, 32:928-961, 2004.

J. Fan and J. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society, Series B*, 62:303-322, 2000.

T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55:757-796, 1993.

D. R. Hunter and R. Li. Variable selection using MM algorithms. *Annals of Statistics*, 33:1617-1642, 2005.

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

J. Huang, J. L. Horowitz, and F. Wei. Variable selection in nonparametric additive models. *Annals of Statistics*, 38:2282-2313, 2010.

J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89:111-128, 2002.

J. Z. Huang, C. O. Wu, and L. Zhou. Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14:763-788, 2004.

Y. Kim, H. Choi, and H. Oh. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103:1665-1673, 2008.

Y. Liu, X. Shen, and W. Wong. Computational development of *psi*-learning. *Proc SIAM 2005 Int. Data Mining Conf.*, 1-12, 2005.

A. Qu, and R. Li. Quadratic inference functions for varying-coefficient models with longitudinal data. *Biometrics*, 62:379-391, 2006.

J. O. Ramsay, and B. W. Silverman. *Functional Data Analysis.* Springer-Verlag: New York, 1997.

X. Shen, W. Pan, Y. Zhu. Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association*, 107:223-232, 2012.

X. Shen, G. C. Tseng, X. Zhang, and W. H. Wong. On ψ-learning. *Journal of the American Statistical Association*, 98:724-734, 2003.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267-288, 1996.

S. Van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36:614-645, 2008.

H. Wang, and Y. Xia. Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, 104:747-757, 2009.

L. Wang, H. Li, and J. Z. Huang. Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103:1556-1569, 2008.

F. Wei, J. Huang, and H. Li. Variable selection and estimation in high-dimensional varying coefficient models. *Statistica Sinica*, 21:1515-1540, 2011.

C. O. Wu, and C. Chiang. Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Statistica Sinica*, 10:433-456, 2000.

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

M. Yuan, and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49-67, 2006.

C. H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894-942, 2010.

P. Zhao, and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541-2563, 2006.

S. Zhou, X. Shen, and D. A. Wolfe. Local asymptotics for regression splines and confidence regions. *Annals of Statistics*, 26:1760-1782, 1998.

H. Zou, and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509-1533, 2008.