DISSERTATION

A PENALIZED ESTIMATION PROCEDURE FOR VARYING COEFFICIENT MODELS

Submitted by

Yan (Catherine) Tu

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2015

Doctoral Committee:

Advisor: Haonan Wang

F. Jay Breidt
Phillip Chapman
Rockey J. Luo

ABSTRACT

A PENALIZED ESTIMATION PROCEDURE FOR VARYING COEFFICIENT MODELS

Varying coefficient models are widely used for analyzing longitudinal data. Various methods for estimating coefficient functions have been developed over the years. We revisit the problem under the theme of functional sparsity. The problem of sparsity, including global sparsity and local sparsity, is a recurrent topic in nonparametric function estimation. A function has global sparsity if it is zero over the entire domain, and it indicates that the corresponding covariate is irrelevant to the response variable. A function has local sparsity if it is nonzero but remains zero for a set of intervals, and it identifies an inactive period of the corresponding covariate. Each type of sparsity has been addressed in the literature using the idea of regularization to improve estimation as well as interpretability. In this dissertation, a penalized estimation procedure has been developed to achieve functional sparsity, that is, simultaneously addressing both types of sparsity in a unified framework. We exploit the property of B-spline approximation and group bridge penalization. Our method is illustrated in simulation study and real data analysis, and outperforms the existing methods in identifying both local sparsity and global sparsity. Asymptotic properties of estimation consistency and sparsistency of the proposed method are established. The term of sparsistency refers to the property that the functional sparsity can be consistently detected.

ACKNOWLEDGEMENTS

First, I would like to thank my parents, He Tu and Lihong Wu, for their selfless love and support throughout my life.

I would also like to thank my advisor Haonan Wang for keep helping and encouraging me over the years I spent abroad. I would also like to thank my friends, the faculty and the staff in the Statistics Department at Colorado State University for their help and support in the past few years.

Last but not least, I would like to thank Dr. Juhyun Park and Dr. Dong Song for their important and insightful comments on my dissertation. They also are both great co-authors to work with.

DEDICATION

To my motherland.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Varying Coefficient Model and Its Interpretation

In today's scientific studies, lots of observations are taken at different time points for each subject, and the information collected in this way are called longitudinal data. To find the linear relationship between response and predictors in longitudinal data, statisticians introduced in varying coefficient model as an extension to linear regression model. Varying coefficient model offers more flexibility in terms of allowing the coefficients varying over time for the predictors. In other words, the impacts of predictors on response are no longer fixed but changeable over observation time. The traditional multiple linear regression model is defined as

$$y = \boldsymbol{x}^T\boldsymbol{\beta} + \epsilon, \tag{1.1}$$

where y is the response, $\boldsymbol{x}$ is the vector of predictors and ϵ is the random error, while the varying coefficient model is defined as

$$y(t) = \boldsymbol{x}^T(t)\boldsymbol{\beta}(t) + \epsilon(t), \tag{1.2}$$

where $y(t)$ is the response at time t, $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))^T$ is the vector of predictors at time t, $\epsilon(t)$ is an error process independent of $\boldsymbol{x}(t)$ and $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ is a vector of time varying regression coefficient functions. This model assumes a linear relationship between the response and predictors at each observation time point but allows the coefficients to vary over time, thus greatly enhances the utility of the standard linear model formulation. Additional flexibility can be gained by employing nonparametric approaches to estimating the coefficient functions, as a parametric approach is limited to capturing the covariate effects in a pre-specified class of functions and is thus prone to model misspecification error. A potential difficulty in using nonparametric regression models however is the issue of interpretability. Although the issue has always drawn attention from statisticians

and scientists alike, traditionally this was considered as a separate issue from estimation, and is often used in practice as a ground for preferring parametric model specifications.

Interpretability in a narrow sense only concerns how to describe the effect of a particular covariate from the fitted model. For a nonparametric component, the effect is expressed as a complex functional form, which does not render an easily recognizable pattern. In this case, some constraints in the functional form would be helpful. Interpretability in a broader sense includes the problem of model selection, as we seek a parsimonious description of the data for a simpler and better interpretable solution. With increasing number of measurable variables available, this becomes more relevant, and can be helped by many new developments in the area of variable selection. These new developments aim at obtaining *sparse solutions* and are applicable to both parametric and nonparametric regression models.

These notions of interpretability were informally used in a rather separate context of the analysis. We argue that both notions of interpretability could be used more formally in relation to the sparsity of the estimates. To distinguish between those two cases, we call the former *local sparsity* and the latter *global sparsity*, and both constitute *functional sparsity* [Tu et al., 2012, Wang and Kai, 2015]. Our purpose is to take both types of sparsity into consideration for model fitting. In particular we focus on situations where (i) some of the predictors do not contribute to forming the relationship and (ii) those contributing predictors are not necessarily active in the whole period. Our estimation procedure is adapted to these cases, producing consistent estimates under these scenarios. The idea of using constrained estimation in this context is not new, however, most available methods address one or the other, but not both. Our contribution is to provide a unified framework for constrained estimation to achieve functional sparsity for varying coefficient linear models.

1.2 Review of Existing Variable Selection Methods

The global sparsity is widely discussed under the realm of variable selection in multiple linear regression. Akaike's information criterion (AIC, Akaike [1973]), Bayesian information criterian (BIC, Schwarz [1978]), and extended BIC (EBIC, Chen and Chen [2008]) are

methods dealing with the trade-off between the goodness of fit and complexity of the model in order to achieve variable selection. The regularized regression approaches opened a new chapter in variable selection, such as ridge regression [Hoerl and Kennard, 1970], lasso [Tibshirani, 1996], and SCAD [Fan and Li, 2001]. The success of lasso has brought about many variants of regularized regression approaches, such as bridge regression [Frank and Friedman, 1993]. The implement of grouping methods makes the existing variable selection methods more efficient while treating models with large number of parameters, such as group lasso [Yuan and Lin, 2006] and group bridge [Huang et al., 2009]. For nonparametric regression, Lin and Zhang [2006] and Ravikumar et al. [2009] extended the idea of lasso to variable selection for additive regression models. An extension to varying coefficient linear models is found in Wang et al. [2008], who proposed a regularized approach based on SCAD penalty.

The ridge regression penalty is defined as

$$p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2^2 = \lambda\sum_{i=1}^{p}\beta_i^2.$$

Since it is based on the L_2 norm of the vector of parameters, it is also considered as an L_2-penalty. Then the ridge regression estimate $\hat{\boldsymbol{\beta}}_{ridge}$ for linear model (1.1) is given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{ridge} &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\boldsymbol{\beta}^T\boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}.\end{aligned}$$

The ridge regression criterion is equivalent to minimizing the squared error loss function subjecting to $\|\boldsymbol{\beta}\|_2^2 \le \theta$. Although biases are introduced while shrinking the least squares estimator, the solution is more robust when $\boldsymbol{X}^T\boldsymbol{X}$ is singular.

The lasso penalty is an L_1-penalty, since it is defined as

$$p_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1 = \lambda\sum_{i=1}^{p}|\beta_i|.$$

Then the lasso estimate $\hat{\boldsymbol{\beta}}_{lasso}$ for linear model (1.1) is given by

$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

The lasso penalized criterion is equivalent to minimizing the squared error loss function subjecting to $\|\boldsymbol{\beta}\|_1 \leq \theta$. Its solutions can be efficiently solved by the lars algorithm [Efron et al., 2004].

The bridge penalty is a modified lasso penalty which is defined as

$$p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{i=1}^{p} |\beta_i|^\gamma,$$

where the tuning parameter $0 < \gamma < 1$. Hence, the constrained area for bridge regression is $\sum_{i=1}^{p} |\beta_i|^\gamma \leq \theta$. Notice that the bridge penalty is not a convex function. The algorithm to solve the optimization problem of bridge regression is stated in Section 2.3.

Figure 1.1 gives a two-dimensional geometric illustration of the constrained areas for ridge, lasso and bridge ($\gamma = 0.5$) penalty functions with $\theta = 1$. From this figure, we notice that the constrained area is a circle for ridge regression, a square for lasso, and a star for bridge regression. The grey circles in the figure represent contour lines for loss functions. The tangent points for lasso and bridge regression are $(0, 1)$, while, for ridge regression, the tangent point is $(1/\sqrt{5}, 2/\sqrt{5})$. This means that lasso and bridge method can achieve sparsity, i.e., zero estimates, compared to ridge regression. And bridge regression is even more likely to shrink small values to zeros than lasso.

The SCAD penalty function is defined as

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & \text{if } 0 \leq |\beta| \leq \lambda \\ -\frac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} & \text{if } \lambda < |\beta| < a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| \geq a\lambda \end{cases}$$

where a is a tuning parameter, and Fan and Li [2001] suggested to choose $a = 3.7$. Under various model settings, the SCAD penalty has been demonstrated to yield consistent estimates for parameter estimation, which also possess the oracle property [Fan and Li, 2001]. Figure 1.2 gives a graphical illustration of the penalty functions mentioned above. Compared with lasso penalty, SCAD penalty overlaps with lasso penalty when $|\beta| \leq \lambda$, but has less penalty on large values and remains constant eventually.

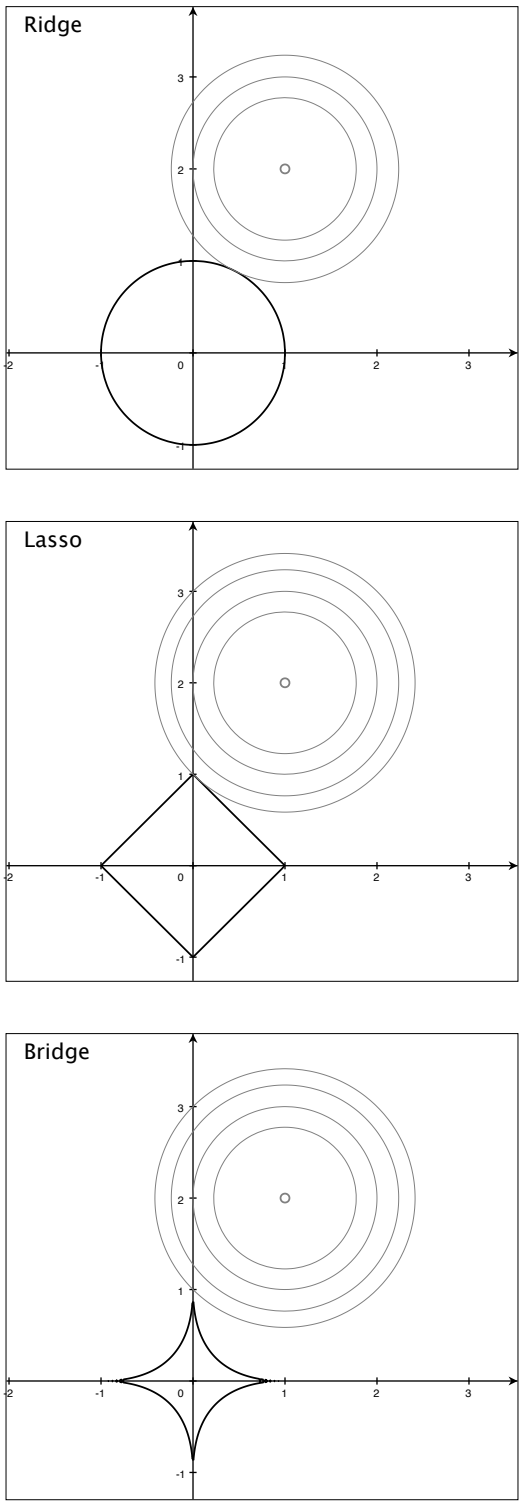

Figure 1.1: Geometric illustration of constrained areas and sparsity for ridge, lasso and bridge methods.

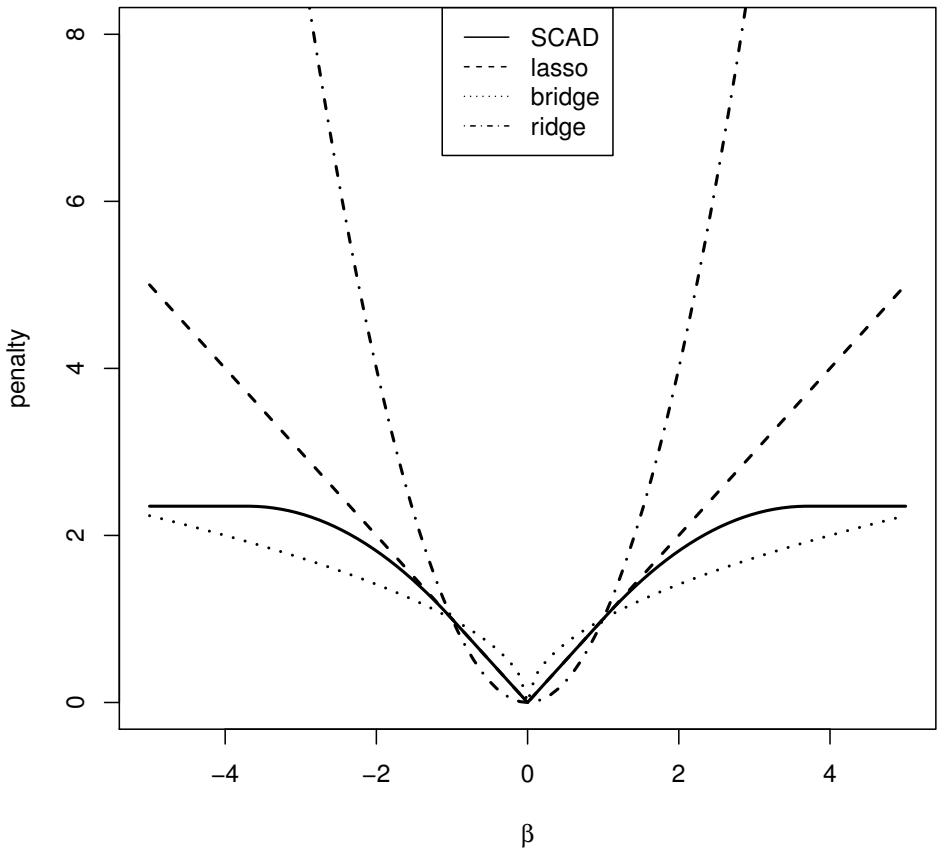

Figure 1.2: Geometric illustration of penalty functions for SCAD (solid line), lasso (dashed line), bridge regression (dotted line), and ridge regression (dot-dashed line).

The group lasso penalty is defined as

$$p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} c_g \|\boldsymbol{\beta}_{A_g}\|_2,$$

where A_g is subset of $\{1, 2, \ldots, p\}$, and $A_g \cap A_l = \emptyset$ for $g \neq l$. Here, $\boldsymbol{\beta}_{A_g}$ is the vector of parameters for the g-th group, and c_g is the corresponding weight. The group lasso penalty can be considered as a combination of L_1-penalty across the groups and L_2-penalty within the groups. Therefore, this method, instead of selecting variables individually, selects the groups of variables. Due to the ridge regression within groups and non-overlapping between groups, group lasso method lacks the ability to achieve sparsity within the groups.

The group bridge method enables the variable selection across groups and within groups. It can be expressed as

$$p_\lambda(\boldsymbol{\beta}) = \lambda \sum_{g=1}^{G} c_g \|\boldsymbol{\beta}_{A_g}\|_1^\gamma,$$

where the groups A_g are allowed to overlap. Compared with group lasso penalty, the grouping method here is more flexible. The group bridge penalty can be considered as a combination of bridge penalty across the groups and L_1-penalty within the groups.

Further, the problem of estimating the models with diverging number of predictors, especially when p is larger than n is another challenge in model selection. Fan et al. [2014] extended an independence screening procedure for linear models [Fan and Lv, 2008] to non-parametric varying coefficient linear models to reduce model complexity. Xue and Qu [2012] also studied the problem of model selection under large-p-small-n setting.

The local sparsity is an emerging issue, mostly attached to nonparametric components. With a purpose of highlighting the effect of the subset of the covariates, current efforts concentrate on simplifying the estimated functional form in order to separate zero estimates from non-zero estimates. Given the underlying continuity assumption of the functions, it is desirable that estimation methods simultaneously determine an active region and an inactive region of the covariates, rather than an active or inactive set of points. Recently James et al. [2009] demonstrated, in the case of functional linear regression with scalar response

variable, that variable selection ideas could be used to achieve this aim. Using B-spline basis representation, they imposed sparsity constraints on the derivatives of the underlying function at a large number of grid points. The procedure uses L_1 penalty on the coefficients through lasso or dantzig selector [Candes and Tao, 2007] for estimation of the coefficients in order to induce exactly zero values in the estimate of the coefficient function for a connected region, and thus remove uncertainty around wiggly fluctuation around zero. It turns out that, due to the overlapping contribution of each coefficient to neighboring regions, independent shrinkage of the coefficients does not necessarily induce zero values in the coefficient function in general, and thus the procedure tends to over-penalize to compensate for indirect control of sparsity. A remedy has been suggested by Zhou et al. [2013] as a two-step estimation procedure with an explicit control of zero and non-zero regions, at the expense of simplicity of the original formulation.

In an attempt to achieve functional sparsity, Tu et al. [2012] considered an alternative regularization method in the context of functional dynamic models. The method proposed is successful in identifying global sparsity but not local sparsity due to the user-defined grouping structure in function approximation. Wang and Kai [2015] focused on local sparsity for single covariate by introducing a new grouping structure.

1.3 An Overview of Our Approaches

In this dissertation, we consider the problem of sparse function estimation under more general setting with multiple covariates. Although detecting sparsity can easily appeal to our intuition, it turns out that this is very difficult to properly formulate in the context of nonparametric function estimation. The major challenge lies in the fact that, for parametric sparsity, an underlying sparse vector is specified whereas for functional sparsity its *true* sparse representation may not be well defined in their respective linear approximation. In fact, the notion of global sparsity can be formulated as variable selection in nonparametric additive models, whose solution exhibits an analogy to regularized variable selection problems in

high dimensional linear regression models; while, for local sparsity, such connection is non-trivial. Most methods proposed in the literature to achieve local sparsity are more in line with (multivariate) parametric methods, likewise, the assumptions and proofs rely on the parametric property. In comparison, our assumptions of smoothness of the functions and proofs are more standard in nonparametric regression and exploit the functional property in more natural manner. Hence, our contribution is to bridge the gap between parametric variable selection and nonparametric functional sparsity in a coherent manner. We also extend our results to high dimensional case. Our theoretical results are extended from fixed dimension p to diverging p. When facing $p > n$, we incorporate the nonparametric independence screening procedure [Fan et al., 2014] to reduce the model complexity and then proceed with our proposed method to estimate relevant coefficient functions.

Our formulation is given in Chapter 2. Our approach is a one-step procedure, and allows us to directly control functional sparsity through the coefficient functions themselves, rather than through their derivatives, hence removing the ambiguity of defining a sparse solution in the latter. In Chapter 3, we study large sample properties of the proposed method and establish consistency and convergence rates of function estimations. Chapter 4 describes simulation studies under different scenarios and two examples of real data analysis are given in Chapter 5. Main proofs are given in Chapter 6. The additional simulation results are provided in Chapter 7.

This dissertation is based on a submitted paper [Tu et al., 2015].

CHAPTER 2

METHODOLOGY

Suppose that, for n randomly selected subjects, observations of the kth subject are obtained at $\{t_{kl}, l = 1, \ldots, n_k\}$ and the measurements satisfy the varying coefficient linear model relationship in (1.2)

$$y_k(t_{kl}) = \boldsymbol{x}_k^T(t_{kl})\boldsymbol{\beta}(t_{kl}) + \epsilon_k(t_{kl}), \tag{2.1}$$

where $\boldsymbol{x}_k(t_{kl}) = (x_1(t_{kl}), \ldots, x_p(t_{kl}))^T$ and $y_k(t_{kl})$ is the response of the kth subject at t_{kl}. We assume that $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ where $\beta_i(t), i = 1, \ldots, p$ are smooth coefficient functions with bounded second derivatives for $t \in \mathcal{T}$. We use spline approximations to represent $\boldsymbol{\beta}(t)$ and formulate a constrained optimization problem for parameter estimation.

2.1 Least squares estimation under B-spline approximation

B-spline approximation has been widely used for estimating smooth nonparametric functions. For detailed discussion about B-splines, see de Boor [2001] and Schumaker [1981]. Specifically, for a smooth function $\beta(t)$, $t \in [0, 1]$, its approximant can be written as

$$\widetilde{\beta}(t) = \sum_{j=1}^{J} \alpha_j B_j(t), \tag{2.2}$$

where $\{B_j(\cdot), j = 1, \ldots, J\}$ is a group of B-spline basis functions of degree $d \geq 1$ and knots $0 = \eta_0 < \eta_1 < \ldots < \eta_K < \eta_{K+1} = 1$. Notice that K is the number of interior knots and $J = K + d + 1$. Here we adopt the definition of B-spline as stated in Definition 4.12 of Schumaker [1981]. In general, performance of B-spline approximation has been well studied. For instance, under some mild conditions, there exists a function $\widetilde{\beta}(t)$ of the form (2.2) such that the approximation error goes to zero. See Theorem 6.27 of Schumaker [1981] for more details.

We write the B-spline approximation for each smooth nonparametric coefficient function as

$$\widetilde{\beta}_i(t) = \sum_{j=1}^{J_i} \alpha_{ij} B_{ij}(t) = \boldsymbol{B}_i(t)^T \boldsymbol{\alpha}_i, \quad t \in [0,1], \quad i = 1, \ldots, p, \tag{2.3}$$

where $\boldsymbol{B}_i(t) = (B_{i1}(t), \ldots, B_{iJ_i}(t))^T$, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ij_i})^T$ and $J_i = K_i + d + 1$. Here K_i is the number of interior knots for $\widetilde{\beta}_i(t)$ which may vary over i. For simplicity, we assume that the knots are evenly distributed over $[0,1]$. Define a block diagonal matrix $\mathcal{B}(t)$ as

$$\mathcal{B}(t) = \text{diag}\{\boldsymbol{B}_1^T(t), \ldots, \boldsymbol{B}_p^T(t)\}.$$

Using (2.3) in the varying coefficient model (2.1) leads to

$$y_k(t_{kl}) \approx \boldsymbol{x}_k^T(t_{kl})\mathcal{B}(t_{kl})\boldsymbol{\alpha} + \epsilon_k(t_{kl}) = \mathcal{U}_k(t_{kl})\boldsymbol{\alpha} + \epsilon_k(t_{kl})$$

where $\mathcal{U}_k(t_{kl}) = \boldsymbol{x}_k^T(t_{kl})\mathcal{B}(t_{kl})$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T)^T$. The least squares criterion of $\boldsymbol{\alpha}$ [Huang et al., 2002] is defined as

$$\ell(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \omega_k \|\boldsymbol{y}_k - \boldsymbol{U}_k\boldsymbol{\alpha}\|_2^2$$

where $\boldsymbol{y}_k = (y_k(t_{k1}), \ldots, y_k(t_{kn_k}))^T$ and $\boldsymbol{U}_k = (\mathcal{U}_k^T(t_{k1}), \ldots, \mathcal{U}_k^T(t_{kn_k}))^T$. Weights ω_k, $k = 1, \ldots, n$, are usually chosen as $\omega_k \equiv 1$ or $\omega_k \equiv 1/n_k$ [Huang et al., 2004]. In this dissertation, for simplicity, we set equal weights to every subject, i.e., $\omega_k \equiv 1$. Putting $\boldsymbol{U} = (\boldsymbol{U}_1^T, \ldots, \boldsymbol{U}_n^T)^T$ and $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, the least squares criterion $l(\boldsymbol{\alpha})$ can be written in matrix form; that is, $l(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2$. Huang et al. [2004] proved that, under certain assumptions, the matrix $\boldsymbol{U}^T\boldsymbol{U}$ is invertible for fixed p. We extended this result for diverging p in Chapter 7. Consequently, $l(\boldsymbol{\alpha})$ has a unique minimizer

$$\widehat{\boldsymbol{\alpha}}_{LSE} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{y},$$

which is the least squares estimator of $\boldsymbol{\alpha}$, and thus, the least squares estimators of coefficient functions are

$$\widehat{\beta}_i^{LSE}(t) = \sum_{j=1}^{J_i} \widehat{\alpha}_{ij}^{LSE} B_{ij}(t), \quad i = 1, \ldots, p,$$

where $\widehat{\alpha}_{ij}^{LSE}$'s are entires of $\widehat{\boldsymbol{\alpha}}_{LSE}$.

2.2 B-spline approximation and Sparsity

From the B-spline approximation theory, there exists a function of the form (2.2) which is very close to the true underlying function. This function is not capable of characterizing functional sparsity of the true function. Here the term "functional sparsity" is a generalization of the "parameter sparsity" in regression models; see Wang and Kai [2015] for more details.

For better illustration, we consider a toy example in Figure 2.1. Here, in the top panel, a smooth function $\beta(t)$ (thick line) with two spline estimates (dashed, dottted) are depicted. In the bottom, a family of cubic B-spline basis functions with 9 interior knots is shown. The "best" fitted function from the L_2 criterion is shown as the dashed line in the upper panel, which signifies a good performance of the approximation. We further note that $\beta(t)$ is zero on $[0, 0.1]$ and $[0.9, 1]$; while, its approximation is not zero except for some singletons. From that aspect, this approximation does not capture the sparsity of the true underlying function. In contrast, the dotted curve depicted in the upper panel, also a linear combination of the B-spline basis functions, *automatically corrects* the function to preserve local sparsity with almost indistinguishable performance.

Note that the least squares method produces consistent function estimates for coefficient functions. Motivated by above discussion on functional sparsity, we develop a new procedure that equips the least squares criterion with a regularization term. Usually, the regularization on parameters is expressed in terms of penalty function. Below we introduce a composite penalty based on the B-spline approximation of the coefficient functions.

2.3 Penalized Least Squares Estimation with Composite Penalty

It is not too difficult to see that global sparsity corresponds to group variable selection of $\boldsymbol{\alpha}_i$ as a whole. To achieve local sparsity, these estimates need to be adjusted in such a way that some of the estimates could be exactly zero. As demonstrated in Section 2.2, we notice that for B-spline approximation, when $\alpha_j = 0$ for $j = l, \ldots, l + d$, the approximation $\widetilde{\beta}(t) = 0$ on the interval $[\eta_{l-1}, \eta_l)$, and especially, when $\alpha_j = 0$ for all j, $\widetilde{\beta}(t) = 0$ over the

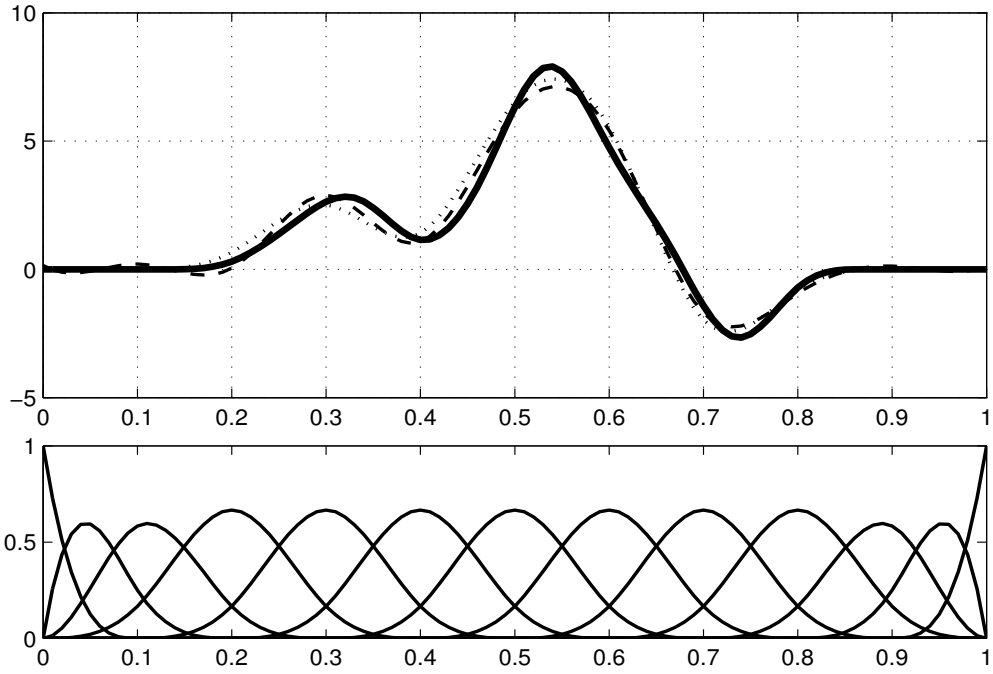

Figure 2.1: Top: a graphical display of a smooth function (solid thick line type) and two approximating functions from a family of cubic B-spline basis functions with 9 equally-spaced interior knots. Bottom: a graphical display of the set of B-spline functions used in the approximation.

entire domain of $[0, M]$. This suggests local sparsity needs to be imposed at the level of a group of neighboring coefficients. To incorporate global sparsity in varying coefficient model, there needs another layer of group structure. These considerations lead us to a composite penalty defined as

$$L_1^\gamma(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \sum_{m=1}^{K_i+1} \left(\sum_{j=m}^{m+d} |\alpha_{ij}| \right)^\gamma,$$

which can be simply written as

$$L_1^\gamma(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \sum_{g=1}^{G_i} \|\boldsymbol{\alpha}_{A_{ig}}\|_1^\gamma, \tag{2.4}$$

where $\boldsymbol{\alpha}_{A_{ig}} = (\alpha_{ig}, \ldots, \alpha_{i(g+d)})'$, $i = 1, \ldots, p$, $g = 1, \ldots, G_i$. The number of groups for the ith coefficient function is $G_i = K_i + 1$.

Equipping the least squares criterion with penalty (2.4), we obtain the penalized least squares (PLS) criterion

$$\mathrm{pl}(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^{p} \sum_{g=1}^{G_i} \|\boldsymbol{\alpha}_{A_{ig}}\|_1^\gamma, \tag{2.5}$$

where $\lambda > 0$ and $0 < \gamma < 1$ are tuning parameters. The proposed penalized least squares estimator (PLSE) $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\lambda, \gamma)$ is defined to be the minimizer of $\mathrm{pl}(\boldsymbol{\alpha})$. Consequently, the functional estimate of $\beta_i(t)$ is given by $\widehat{\beta}_i(t) = \boldsymbol{B}_i(t)^T \widehat{\boldsymbol{\alpha}}_i$, where $\widehat{\boldsymbol{\alpha}}_i$ is the subvector of $\widehat{\boldsymbol{\alpha}}$.

For $\gamma \in (0, 1)$, the penalized criterion $\mathrm{pl}(\boldsymbol{\alpha})$ is not a convex function of $\boldsymbol{\alpha}$. We implement the iterative algorithm proposed and studied by Huang et al. [2009] to minimize (2.5). The algorithm is outlined as follows.

Step 1. Obtain an initial value $\boldsymbol{\alpha}^{(0)}$.

Step 2. For a given tuning parameter λ_n, and for $l = 1, 2, \ldots$, compute

$$\theta_{ig}^{(l)} = \left(\frac{1-\gamma}{\tau_n \gamma} \right)^\gamma \|\boldsymbol{\alpha}_{A_{ig}}^{(l-1)}\|_1^\gamma, \text{ for } i = 1, \ldots, p, \ g = 1, \ldots, G_i,$$

where $\tau_n = (\lambda_n/n)^{1/(1-\gamma)} \gamma^{\gamma/(1-\gamma)} (1-\gamma)$.

Step 3. Compute

$$\boldsymbol{\alpha}^{(l)} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2 + \sum_{i=1}^{p}\sum_{g=1}^{G_i}(\theta_{ig}^{(l)})^{1-1/\gamma}\|\boldsymbol{\alpha}_{A_{ig}}\|_1.$$

Step 4. Repeat steps 2 and 3 until convergence.

2.4 Variance Estimation

In this section we consider the problem of finding the asymptotic variance of our proposed estimator of the coefficient functions. Let $\widehat{\boldsymbol{\alpha}}_S$ denote the non-zero estimators of the coefficients α_{ij}'s, then by Step 3 in the aforementioned algorithm and the Karush-Kuhn-Tucker condition, we have

$$\widehat{\boldsymbol{\alpha}}_S = \left(\boldsymbol{U}_S^T\boldsymbol{U}_S + \frac{1}{2}\boldsymbol{\Theta}_S\right)^{-1}\boldsymbol{U}_S^T\boldsymbol{y},$$

where $\boldsymbol{U}_S$ is the sub-matrix of $\boldsymbol{U}$ with each column corresponding to the selected α_{ij}, and $\boldsymbol{\Theta}_S$ is a diagonal matrix

$$\text{diag}\left\{\sum_{g:A_{ig}\ni j}\widehat{\theta}_{ig}^{1-1/\gamma}/|\widehat{\alpha}_{ij}|,\ \text{for } \widehat{\alpha}_{ij}\neq 0\right\}.$$

The variance σ^2 can be estimated by $\widehat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2/n$. Thus, similar to Wang et al. [2008], the asymptotic variance of $\widehat{\boldsymbol{\alpha}}_S$ may be expressed as

$$\text{avar}(\widehat{\boldsymbol{\alpha}}_S) = \left(\boldsymbol{U}_S^T\boldsymbol{U}_S + \frac{1}{2}\boldsymbol{\Theta}_S\right)^{-1}\boldsymbol{U}_S^T\boldsymbol{U}_S\left(\boldsymbol{U}_S^T\boldsymbol{U}_S + \frac{1}{2}\boldsymbol{\Theta}_S\right)^{-1}\widehat{\sigma}^2.$$

Let $\mathcal{B}_i(t)$ be the i-th row of the basis matrix $\mathcal{B}(t)$. Thus, the functional estimate of $\beta_i(t)$ can be written as $\widehat{\beta}_i(t) = \mathcal{B}_i(t)\widehat{\boldsymbol{\alpha}}$. Correspondingly, the asymptotic variance of $\widehat{\beta}_i(t)$ is

$$\text{avar}(\widehat{\beta}_i(t)) = \mathcal{B}_{iS}(t)\text{avar}(\widehat{\boldsymbol{\alpha}}_S)\mathcal{B}_{iS}^T(t), \tag{2.6}$$

where $\mathcal{B}_{iS}(t)$ is the sub-vector of $\mathcal{B}_i(t)$ with each element corresponding to the selected α_{ij}. Note that the estimator of $\boldsymbol{\alpha}$ depends on the choice of λ, so the asymptotic variances of $\widehat{\boldsymbol{\alpha}}_S$ and $\widehat{\beta}_i(t)$ are also tuning parameter dependent.

2.5 Choice of Tuning Parameters

To implement the proposed method, it is necessary to select two tuning parameters, $0 < \gamma < 1$ and $\lambda > 0$. The tuning parameter $\lambda > 0$ balances the trade-off between goodness-of-fit and model sparsity. When λ is large, we have strong penalization and thus are more likely to obtain a sparse solution with poor model fitting. On the other hand, with small λ, we would select more variables and get better estimation results but lose control of functional sparsity.

The tuning parameter γ influences the performance of group selection. Too small or too large value of γ could lead to inefficient group variable selection. When γ is close to 1, (2.4) is close to the L_1 penalty. Consequently, the minimizer of (2.5) may not achieve the goal of group variable selection. Small γ will results in large θ_{ig}'s and will yield a sparse solution. In this dissertation, we fix γ to 0.5 for computational purpose [Huang et al., 2009].

In classical nonparametric approaches, the criteria such as AIC, BIC and GCV [Wahba, 1990] are commonly used for choosing λ. It has been noted in previous analyses that the AIC and GCV criteria tend to select more variables, and are better suited for prediction purpose. We use a BIC-type criterion in our analysis reported in Chapter 4. However, when comparing models with high dimensions, and the number of parameters also diverges with sample size n, the ordinary BIC may not be as efficient as usual. Then we use the extended BIC (EBIC) [Huang et al., 2010] to determine the choice of λ. The EBIC is given by

$$EBIC(\lambda) = \log\left(\|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}(\lambda)\|_2^2/N\right) + K(\lambda)\log(N)/N + \nu K(\lambda)\log\left(\sum_{i=1}^{p} J_i\right)/N,$$

where $N = \sum_{k=1}^{n} n_k$, $\widehat{\boldsymbol{\alpha}}(\lambda)$ is the penalized estimator of $\boldsymbol{\alpha}$ given λ, and $K(\lambda)$ is the total number of non-zero estimates in $\widehat{\boldsymbol{\alpha}}(\lambda)$. Moreover, $0 \leq \nu \leq 1$ is a constant and $\sum_{i=1}^{p} J_i = \sum_{i=1}^{p}(K_i + d + 1)$ is the total number of parameters in the full model. Note that when $\nu = 0$ the EBIC is the same as BIC, but when $\nu > 0$, EBIC puts more penalty on overfitting.

2.6 Ultra-High Dimension Problem

We extend our proposed method to models with high dimensional p, in particular, to models with the number of covariates p even larger than the sample size n. For ultra-high dimensional models, it is inefficient to directly apply the interval-based group penalization approach. To estimate large-p small-n models, we combine the nonparametric independence screening method [Fan et al., 2014] with our proposed method. Those authors stated the consistency of the screening step in model selection.

The outline of algorithm including initial screening step is stated as follows:

Step 1. For $i = 1, \ldots, p$, compute

$$u_i = \sum_{k=1}^{n} \sum_{l=1}^{n_k} [(\widehat{\beta}_{0i}^{LSE}(t_{kl}) + x_i(t_{kl})\widehat{\beta}_i^{LSE}(t_{kl}))^2 - (\widehat{\beta}_0^{LSE}(t_{kl}))^2]$$

where $\widehat{\beta}_0^{LSE}$ is estimate of model with only intercept term, and $\widehat{\beta}_{0i}$ and $\widehat{\beta}_i$ are estimates of model with intercept term and the covariate x_i. Sort u_i from highest to lowest and select the variables corresponding to the top M u_i values.

Step 2. Follow the iterative algorithm in Section 2.3 with selected M variables, and find the penalized least squares estimates.

For instance, when we start with a model with $p = 1000$ covariates while only a few of them are important, in the screening step, $M = 8$ covariates are selected, and the original ultra-high dimensional model is reduced to a model with 8 covariates. Afterwards, the penalized method is applied to the reduced model for further investigation.

CHAPTER 3

Large Sample Properties

We study large sample properties of our proposed penalized least squares estimator $\widehat{\beta}_i(t)$, $i = 1, \ldots, p$, when the number of sampled subjects n goes to infinity. We assume in the proofs that the number of observations for each subject n_k is bounded but a similar argument can be applied to the case when n_k increases to infinity with n [Huang et al., 2004]. The number of interior knots increases with n, so we write $K_i = K_{in}$ for each $i = 1, \ldots, p$, and denote $K_n = \max_{0 \le i \le p} K_i$. The standard regularity conditions for varying coefficient linear models [Huang et al., 2004, Wang et al., 2008] are given in Chapter 6.

It is known that, by Theorem 6.27 of Schumaker [1981], any smooth coefficient function $\beta_i(t)$ with bounded second derivative has a B-spline approximant $\widetilde{\beta}_i(t)$ of form (2.3) and the approximation error is of order $O(K_{in}^{-2})$. Denote its sparse modification introduced in Section 2.3 by $\widetilde{\beta}_i^0(t)$ with its coefficients $\widetilde{\boldsymbol{\alpha}}^0$.

For our mathematical convenience, we classify all group indices $\{1, \ldots, G_i\}$ for the coefficient function $\beta_i(t)$ into two groups defined as

$$\begin{aligned} \mathcal{A}_{i1} &= \{g : \max_{t \in [\eta_{g-1}, \eta_g)} \mid \beta_i(t) \mid > C_i K_n^{-2}\}, \\ \mathcal{A}_{i2} &= \{g : 0 \le \max_{t \in [\eta_{g-1}, \eta_g)} \mid \beta_i(t) \mid \le C_i K_n^{-2}\}, \end{aligned}$$

for some positive constant C_i. For sufficiently large C_i, the zero region $\{t : \beta_i(t) = 0\}$ is a subset of $\cup_{g \in \mathcal{A}_{i2}} [\eta_{g-1}, \eta_g)$.

Note that for a vector-valued square integrable function $A(t) = (a_1(t), \ldots, a_m(t))^T$ with $t \in [0, M]$, $\|A\|_2$ denotes the L_2 norm defined by $\|A\|_2 = (\sum_{l=1}^m \|a_l\|_2^2)^{1/2}$ where $\|a_l\|_2$ is the usual L_2 norm in function space.

Now, we establish the consistency of our proposed penalized estimator.

Theorem 1 (Consistency)**.** *Suppose that assumptions (A1)-(A5) in Chapter 6 are satisfied. If* $0 < \gamma < 1$ *and* $K_n = O(n^{1/5})$ *and the following assumption*

(S1) *for* $\widetilde{\boldsymbol{\alpha}}^0$ *defined above,*

$$\lambda_n (d+1)^{1/2} \left(\sum_{i=1}^{p} \sum_{g \in \mathcal{A}_{i1}} \| \widetilde{\boldsymbol{\alpha}}^0_{A_{ig}} \|_1^{2(\gamma-1)} \right)^{1/2} = O(n^{1/2})$$

holds, then we have $\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \|_2 = O_p(n^{-2/5})$*, where* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

Assumption (S1) provides a bound on the rate of λ_n growing with n. The convergence rate established in Theorem 1 is essentially the optimal one [Stone, 1982]. In fact, the result remains valid for more general class of functions, e.g., the collection of functions whose derivatives satisfying the Hölder condition. Next, Theorem 2 states that our proposed penalized method is consistent in detecting functional sparsity. That is if $\beta_i(t) = 0$ for $t \in [\eta_{l-1}, \eta_l)$, then the proposed estimator will produce $\widehat{\boldsymbol{\alpha}}_{A_{il}} = \mathbf{0}$ to identify local sparsity with probability converging to 1. And if $\beta_i(t) = 0$ for all t, then the proposed method will have $\widehat{\boldsymbol{\alpha}}_{A_{il}} = \mathbf{0}$ for all $l = 1, \ldots, K_i + 1$ with probability converging to 1.

Theorem 2 (Sparsistency)**.** *If assumptions in Theorem 1 and the following assumption*

(S2) $\lambda_n K_n^{\gamma-1} n^{-\gamma/2} \longrightarrow \infty$

are satisfied, then we have for every i*,* $i = 1, \ldots, p$*,* $(\widehat{\boldsymbol{\alpha}}_{A_{ig}} : g \in \mathcal{A}_{i2}) = \mathbf{0}$ *with probability converging to 1 as* n *goes to* ∞*.*

It is not surprising that our proposed method will yield a slightly more sparse functional estimate. This is due to the fact that, for all intervals belonging to $\mathcal{A}_{i2}$, the value of $\beta_i(t)$ is quite small, the same order as the optimal rate, and is *indistinguishable* from zeros.

Theorem 3 (Diverging p)**.** *Suppose that assumptions (A1)-(A5) in Chapter 6 are satisfied. In addition, assume that*

(A0) $\quad \lim_n p^2 K_n (\log p + \log K_n)/n = 0\,,$

and the number of relevant covariates $p_0 < p$ *is fixed. If* $0 < \gamma < 1$ *and* $K_n = O(n^{1/5})$ *and the following assumption*

(S1′) $\lambda_n (d+1)^{1/2} \left(\sum_{i=1}^{p} \sum_{g \in \mathcal{A}_{i1}} \| \widetilde{\boldsymbol{\alpha}}^0_{A_{ig}} \|_1^{2(\gamma-1)} \right)^{1/2} = O(p^{1/2} n^{1/2})$

holds, we have $\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \|_2 = O_p(p^{1/2} n^{-2/5})$. *In addition, if*

(S2′) $\lambda_n K_n^{\gamma-1} n^{-\gamma/2} p^{-1+\gamma/2} \to \infty$

is satisfied, the sparsistency result in Theorem 2 holds.

Assumptions (S1′) and (S2′) are in parallel to assumptions (S1) and (S2), but allow p to increase with n. Different from the classic large-p small-n models, the diverging order of p is constrained by (A0) as well as these two assumptions. In fact, for finite p, this assumption can be simplified as $\lim_n K_n \log K_n / n = 0$, which is satisfied when $K_n = O(n^{1/5})$. Moreover, we assume that the number of relevant covariates is a fixed value, independent of p and n, which is comparable to the assumption in Xue and Qu [2012]. This enables us to control overall approximation errors for nonparametric coefficient functions.

CHAPTER 4

Simulation Study

We conducted simulation studies to assess the performance of our proposed method. Relative performance is measured against those from LSE and Lasso methods. Since our penalty function relies on the accuracy of the B-spline representation of the true coefficient functions, we also investigated the impact of the approximation error on the sparsity detection. We conducted Monte Carlo simulations under three different scenarios to study the effect of increasing dimension with $p = 5, 20, 1000$. In each repetition subjects are randomly selected according to the following varying coefficient model specification

$$y_k(t_{kl}) = \sum_{i=1}^{p} x_{ki}(t_{kl})\beta_i(t_{kl}) + \epsilon_k(t_{kl}), \quad l = 1, \ldots, n_k, \quad k = 1, 2, \ldots, n,$$

where $x_1(t)$ is constant 1, $x_i(t)$, $i = 2, 3, 4$ are similar to those considered in Huang et al. [2002]: $x_2(t)$ is a uniform random variable over $[4t, 4t + 2]$; $x_3(t)$ conditioning on $x_2(t)$ is a normal random variable with mean zero and variance $(1 + x_2(t))/(2 + x_2(t))$; and $x_4(t)$, independent of $x_2(t)$ and $x_3(t)$, is Bernoulli(0.6). In Scenario 1, we add a redundant variable $x_5(t)$ from normal distribution with mean zero and variance $0.1\exp(t)$ for illustration of global sparsity. And thus, we have $p = 5$ with $n = 200$ and Scenario 1 represents models with fixed and finite dimension. In Scenario 2, we increase p to 20 and consider two subscenarios with different sample sizes $n = 200$ (Scenario 2.1) and $n = 500$ (Scenario 2.2). The extra predictors with zero coefficient functions are defined as $x_i(t) = Z_i(t) + 3/20 \sum_{l=1}^{5} x_l(t)$ for $i = 6, \ldots, 20$ with $Z_i(t)$'s iid from standard normal distribution. In Scenario 3, p is further increased to 1000 while keeping n fixed to 200 to demonstrate the estimation performance of ultra-high dimensional models. The redundant variables are defined in the similar way as those in Scenario 2.

The number of measurements available varies across the subjects. For each subject a sequence of 40 possible observation time points between 0 and 1 are considered, but each

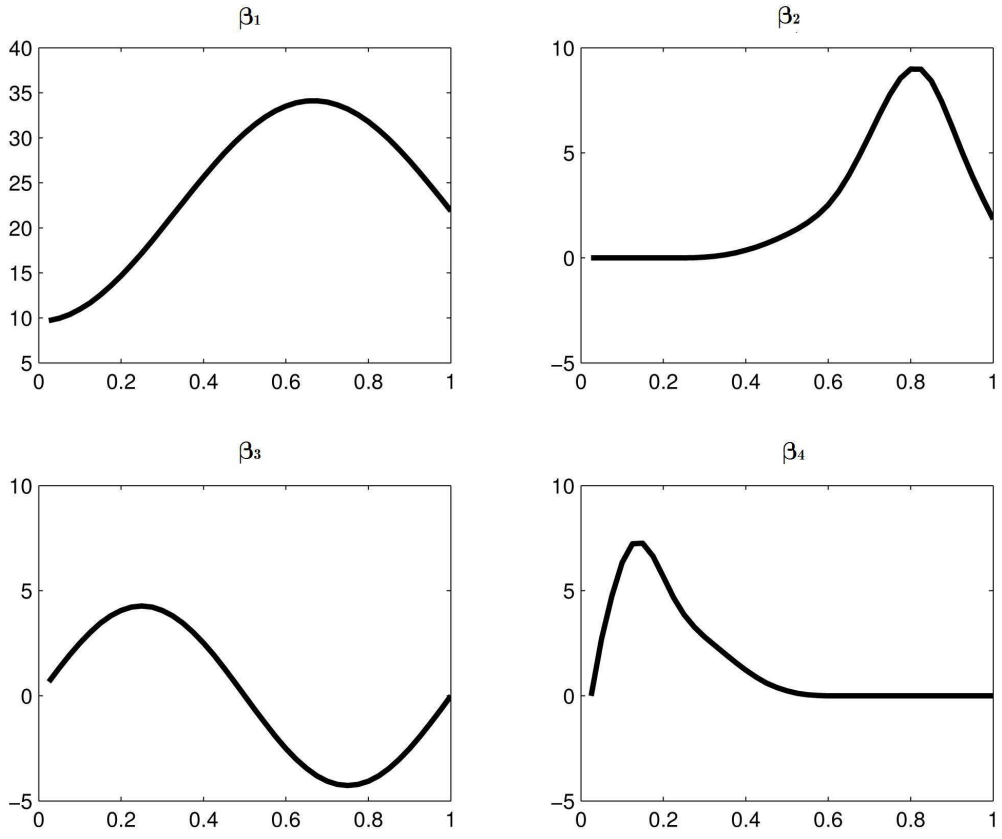

Figure 4.1: A graphical illustration of the coefficient functions β_i $i = 1, \ldots, 4$ in Scenarios 1, 2 and 3.

time point has a chance of 0.4 being selected. The random errors $\epsilon_k(t_{kl})$ are independent of the predictors, and are generated as $\epsilon_k(t) = \epsilon_k^{(1)}(t) + \epsilon_k^{(2)}(t)$ where $\epsilon^{(1)}(t)$ is a Gaussian process with mean zero and covariance function $cov(\epsilon_k^{(1)}(t), \epsilon_k^{(1)}(s)) = \exp(-0.5|t-s|)$ for the same subject k and uncorrelated for different subjects, and $\epsilon_k^{(2)}(t)$'s are iid from normal distribution with mean zero and variance 0.25. The nonzero coefficient functions used in all scenarios are displayed in Figure 4.1. The coefficient functions do not belong to the B-spline function space. We also considered the case where coefficient functions are the elements of the B-spline function space. The additional simulation results are summarized in Chapter 7. In Scenario 1, we use 10-fold cross-validation based on unpenalized LSE to select the number of interior knots for B-spline approximation in each repetition. For models with large p, we fix the number of interiors knots to 11 in Scenarios 2 and 3.

The proposed estimator PLSE_γ is compared with LSE and Lasso estimator in terms of bias and mean integrated squared error (MISE), based on $R = 200$ repetitions, computed as

$$\widehat{Bias}_i(u) = \frac{1}{R}\sum_{r=1}^{R} \widehat{\beta}_i^{(r)}(u) - \beta_i(u), \quad i = 1, \ldots, p,$$

$$\widehat{MISE}_i = \frac{1}{R}\sum_{r=1}^{R} \int_0^1 (\widehat{\beta}_i^{(r)}(u) - \beta_i(u))^2 du, \quad i = 1, \ldots, p.$$

where $\widehat{\beta}_i^{(r)}$ is the estimated coefficient function from the r-th repeated study. In addition, we introduce the following summary measures for comparison of functional sparsity:

(a) C_0: average number of correctly identified constant zero coefficient functions

(b) I_0: average number of incorrectly identified constant zero coefficient functions

(c) $C_{i,0}$: average length of correctly identified zero intervals for the i-th coefficient function

(d) $I_{i,0}$: average length of incorrectly identified zero intervals for the i-th coefficient function.

Note that (a) and (b) summarize global sparsity; while, (c) and (d) summarize local sparsity.

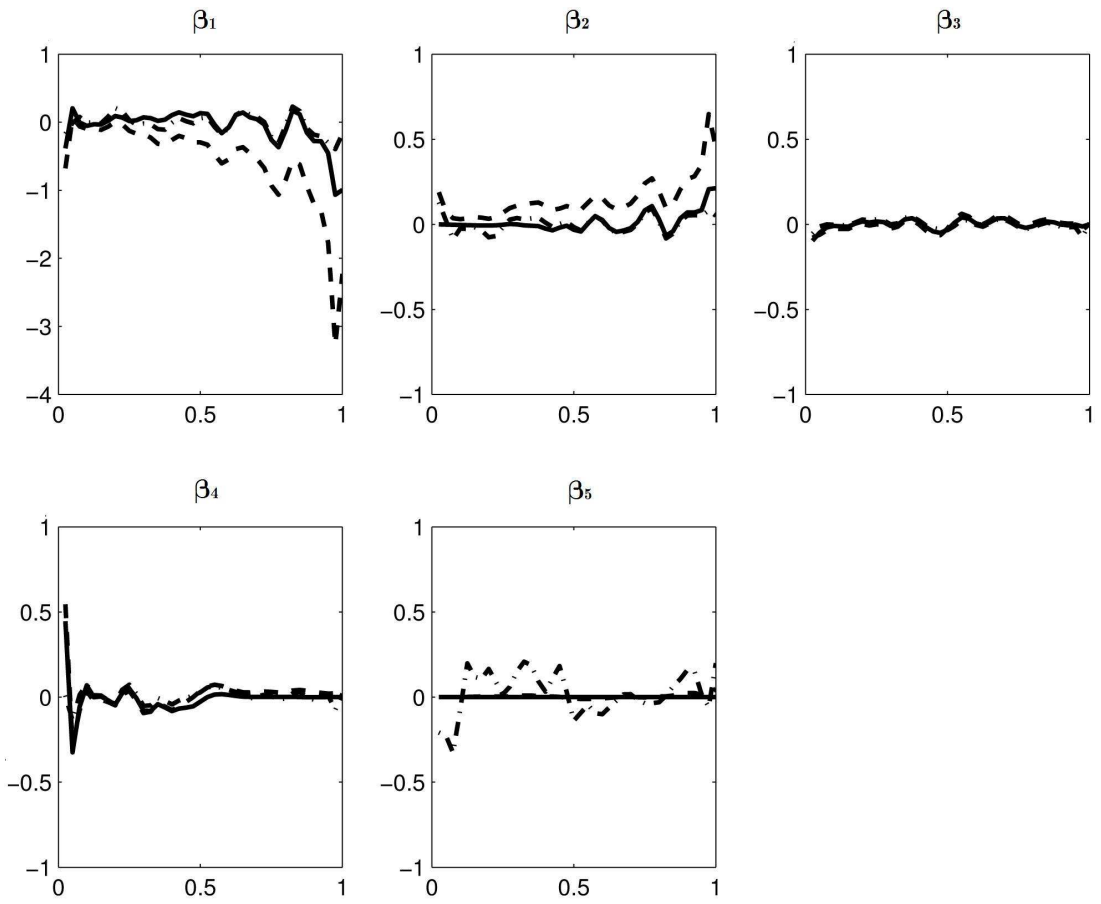

Figure 4.2: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 1.

Table 4.1: Comparison of MISE for each coefficient function in Scenario 1.

Method	MISE				
	β_1	β_2	β_3	β_4	β_5
LSE	0.3644	0.0301	0.0107	0.0359	0.3506
Lasso	1.1223	0.0609	0.0099	0.0326	0.0046
$\text{PLSE}_{0.5}$	0.4015	0.0234	0.0093	0.0262	0

Table 4.2: Sparsity summary measures (a)-(d) in Scenario 1. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 6$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

estimator	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	$C_{5,0}$	$I_{5,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.018	0.001	0	0.010	0.065	0.003	0.838	0	0.41	0
$\text{PLSE}_{0.5}$	0	0	0.182	0.031	0	0.019	0.355	0.031	1	0	1	0
true model	0	1	0.225	0.775	0	1	0.425	0.575	1	0	1	4

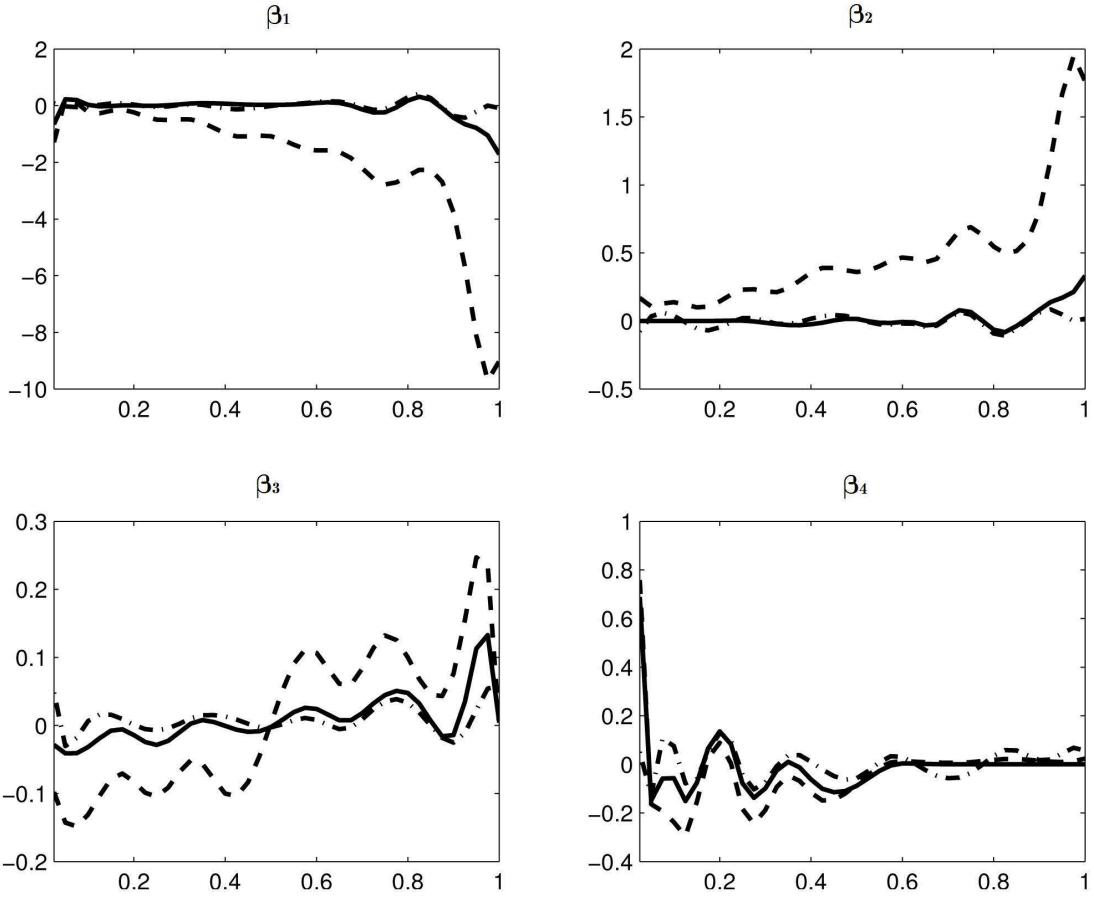

Figure 4.3: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 2.1.

Table 4.3: Comparison of MISE for each coefficient function in Scenario 2.1.

Method	MISE				$\max_{i\geq 5}$ MISE$_i$
	β_1	β_2	β_3	β_4	
LSE	0.2300	0.0199	0.0091	0.0298	0.3104
Lasso	9.4774	0.4487	0.0176	0.0350	0.0009
PLSE$_{0.5}$	0.3282	0.0157	0.0076	0.0252	0.0000

Table 4.4: Sparsity summary measures (a)-(d) in Scenario 2.1. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.005	0	0	0.022	0.255	0.013	2.8	0
PLSE$_{0.5}$	0	0	0.218	0.026	0	0.029	0.392	0.022	15.91	0
true model	0	1	0.225	0.775	0	1	0.425	0.575	16	4

Table 4.5: Comparison of MISE for each coefficient function in Scenario 2.2.

Method	MISE				$\max_{i\geq 5}$ MISE$_i$
	β_1	β_2	β_3	β_4	
LSE	0.0855	0.0075	0.0031	0.0123	0.1127
Lasso	3.2786	0.1549	0.0064	0.0241	0.0003
PLSE$_{0.5}$	0.1036	0.0052	0.0027	0.0139	0.0000

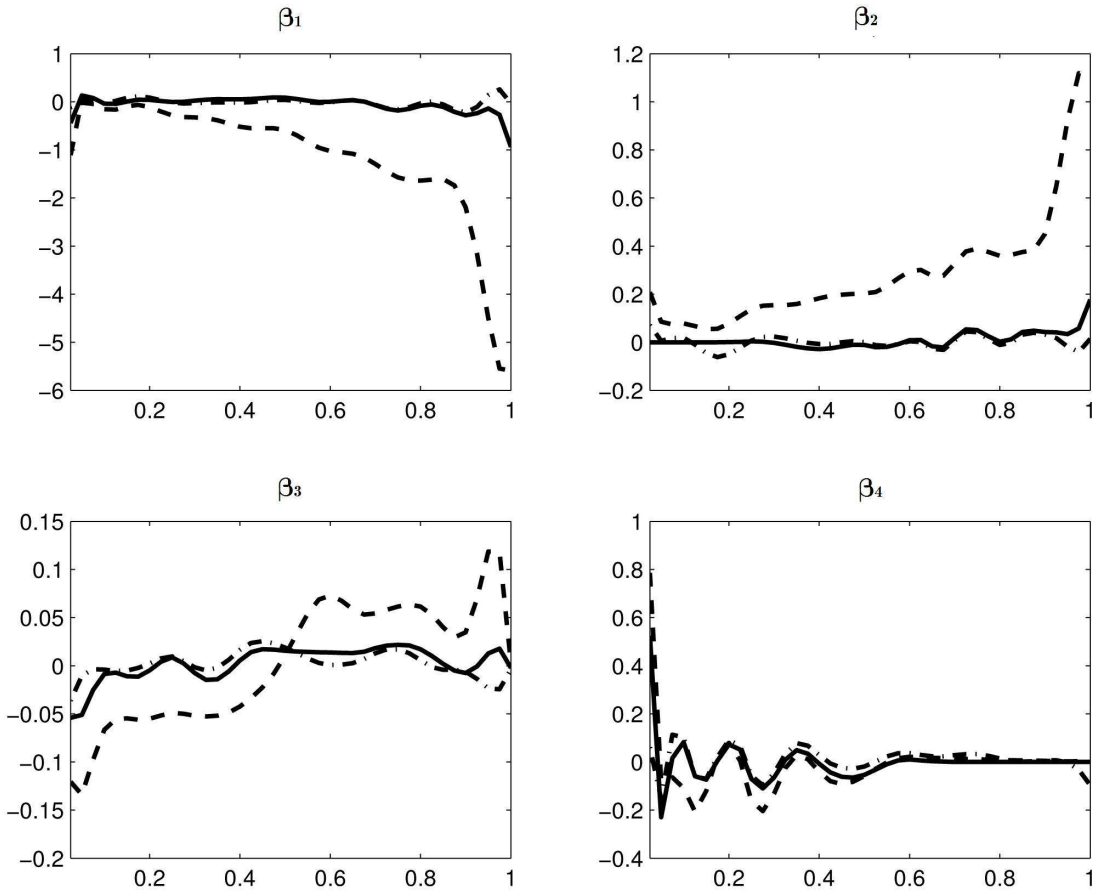

Figure 4.4: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 2.2.

Table 4.6: Sparsity summary measures (a)-(d) in Scenario 2.2. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.004	0	0	0.023	0.243	0.007	2.725	0
$\text{PLSE}_{0.5}$	0	0	0.206	0.019	0	0.022	0.369	0.012	15.895	0
true model	0	1	0.225	0.775	0	1	0.425	0.575	16	4

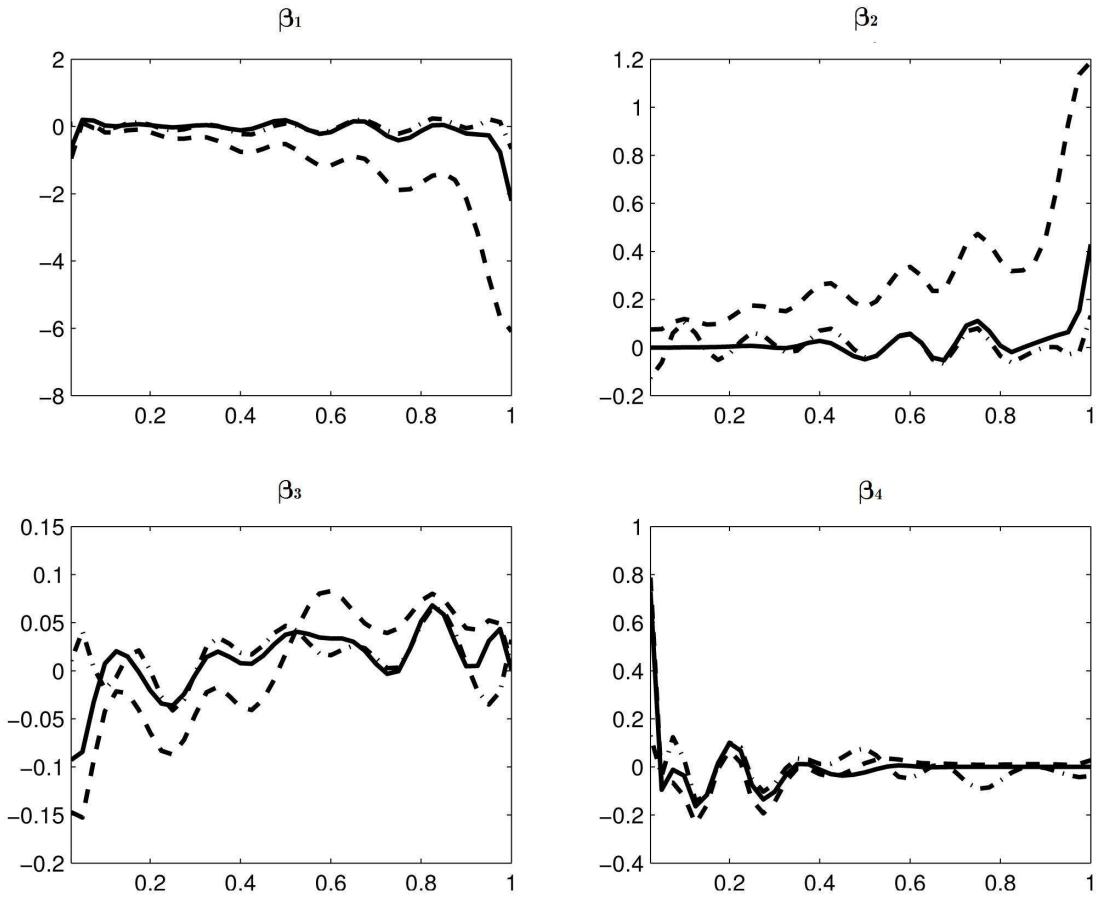

Figure 4.5: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 3.

Table 4.7: Comparison of MISE for each coefficient function in Scenario 3.

Method	MISE				$\max_{i\geq 5} \text{MISE}_i$
	β_1	β_2	β_3	β_4	
LSE	0.2249	0.0199	0.0086	0.0288	0.0080
Lasso	3.6944	0.1793	0.0101	0.0300	0.0014
$\text{PLSE}_{0.5}$	0.3434	0.0167	0.0073	0.0252	0.0001

Table 4.8: Sparsity summary measures (a)-(d) in Scenario 3. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.009	0.001	0	0.02	0.245	0.006	0.045	0
$\text{PLSE}_{0.5}$	0	0	0.211	0.022	0	0.024	0.375	0.013	4.875	0
true model	0	1	0.225	0.775	0	1	0.425	0.575	5	4

Bias of each method is compared in Figures 4.2 - 4.5 and MISE values for every coefficient function are compared in Tables 4.1, 4.3, 4.5 and 4.7. The last column of MISE tables for Scenarios 2 and 3 provides the maximum MISE among the zero coefficient functions, as the selected variables vary from sample to sample. In general the results indicate comparable performances across different scenarios. We also note that $\text{PLSE}_{0.5}$ has zero bias and MISE in estimating the zero coefficient function $\beta_5(\cdot)$, successfully achieving global sparsity, and in particular, has smaller squared error than Lasso method in Scenario 1. And such outstanding performance in global sparsity also appears in high dimension scenarios. Local sparsity is better demonstrated in Tables 4.2, 4.4, 4.6 and 4.8. In summary, the simulation results demonstrate that our proposed method not only has an advantage in achieving local sparsity compared with Lasso and LSE, but also can ensure global sparsity for finite dimensional models. Moreover, this advantage is carried onto models with diverging dimension and models with ultra-high dimension after incorporating a screening step.

In addition, in order to assess the usefulness of the asymptotic formula for the standard errors in (2.6), both asymptotic and empirical standard errors based on 200 repetitions are calculated with fixed number of knots and plotted in Figures 4.6, 4.9, 4.11 and 4.13, which show a good agreement between them. In Scenario 1, we also calculated standard errors with adaptive number of knots shown in Figure 4.7. It can be seen that the variation in number of knots greatly increases the variation in estimation of coefficient functions. The

corresponding empirical 95% confidence intervals (dash-dotted) are shown in Figures 4.8, 4.10, 4.12 and 4.14, with the average estimates (solid line), and the true coefficient functions (dashed line) overlayed.

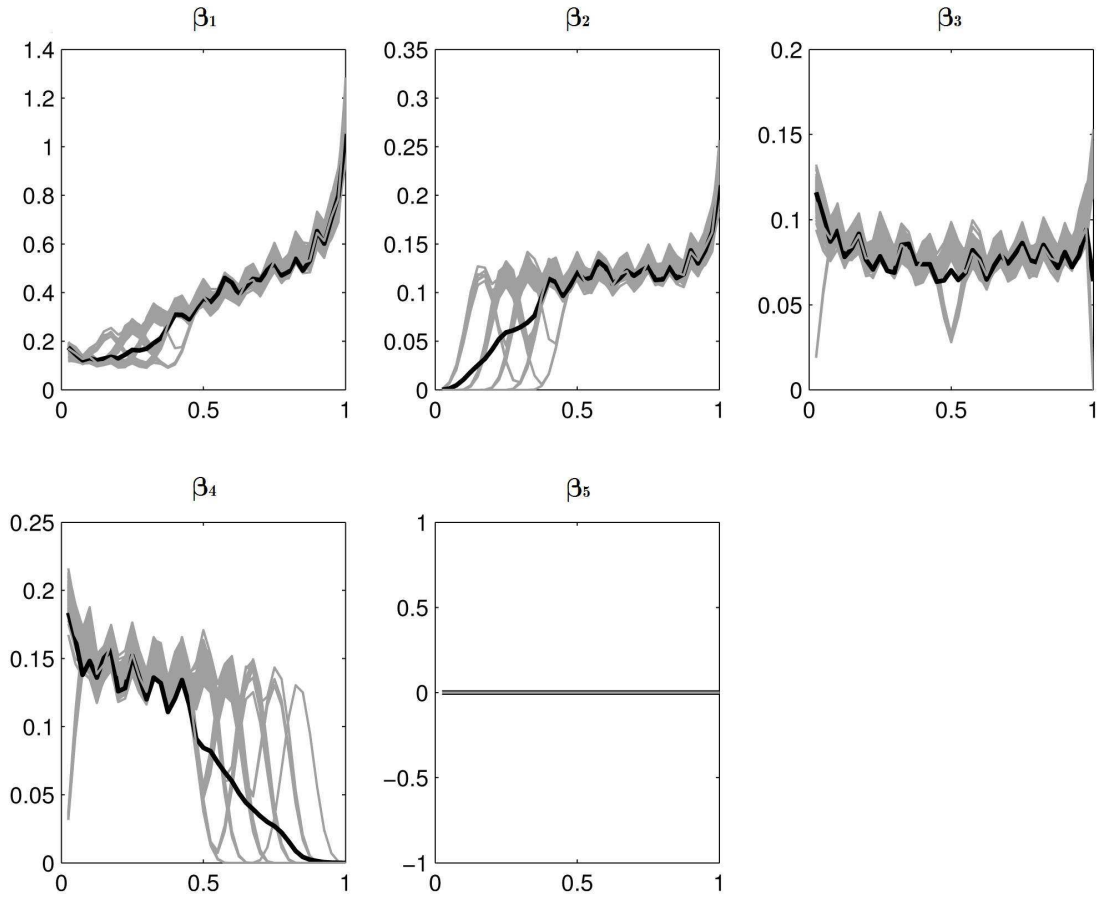

Figure 4.6: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions with fixed number of knots in Scenario 1.

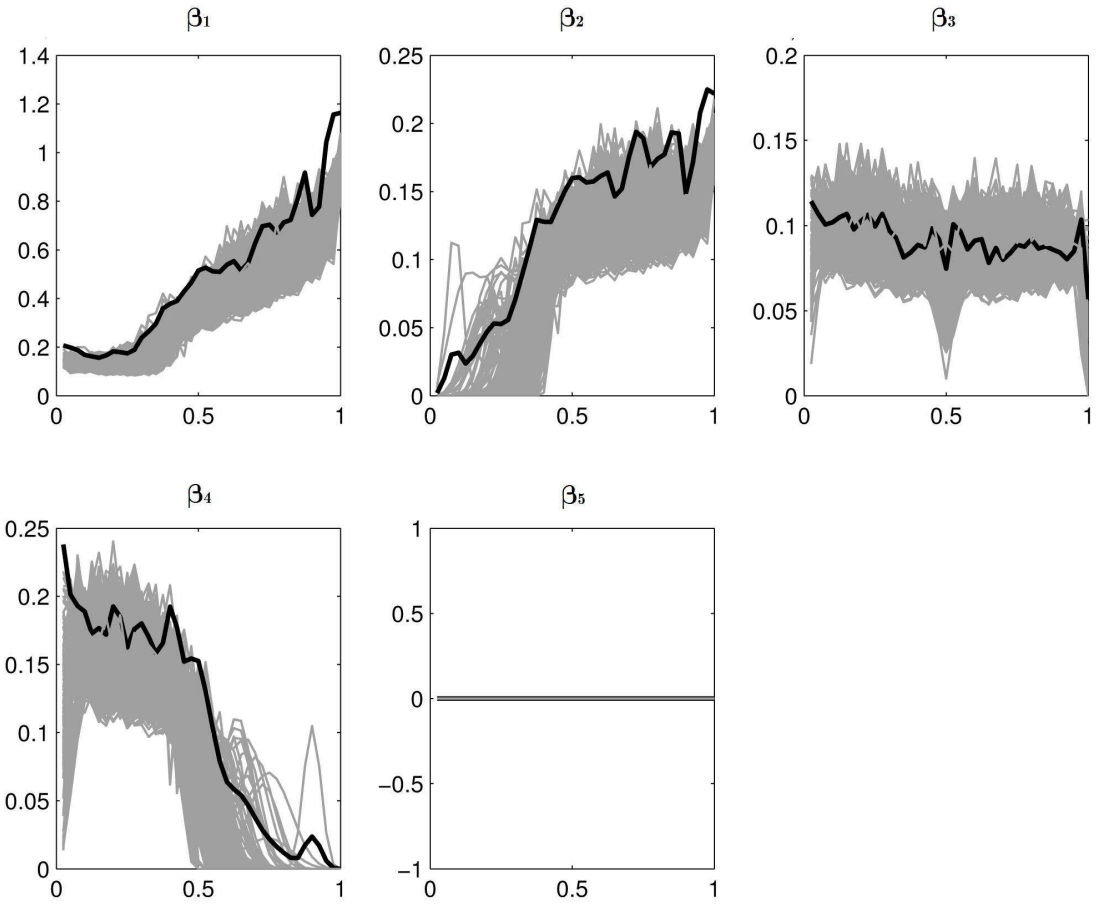

Figure 4.7: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions with adaptive number of knots in Scenario 1.

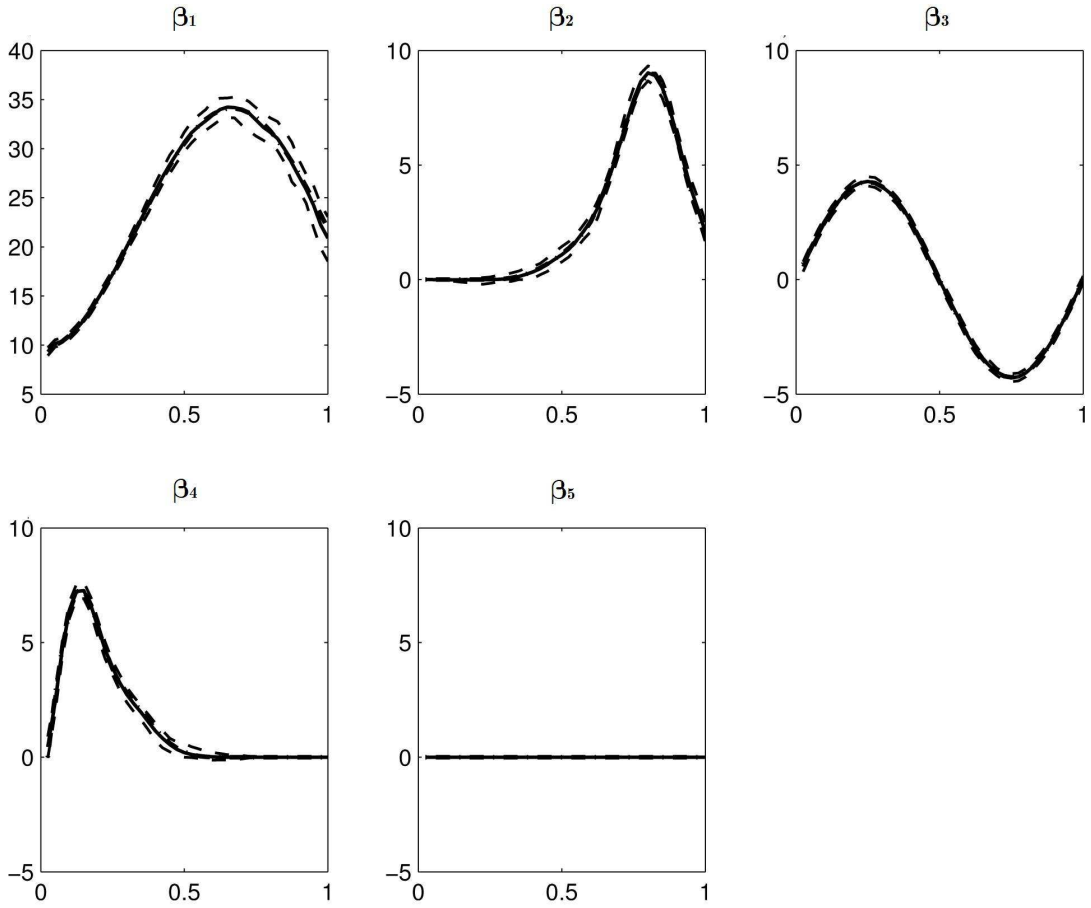

Figure 4.8: Empirical confidence intervals in Scenario 1.

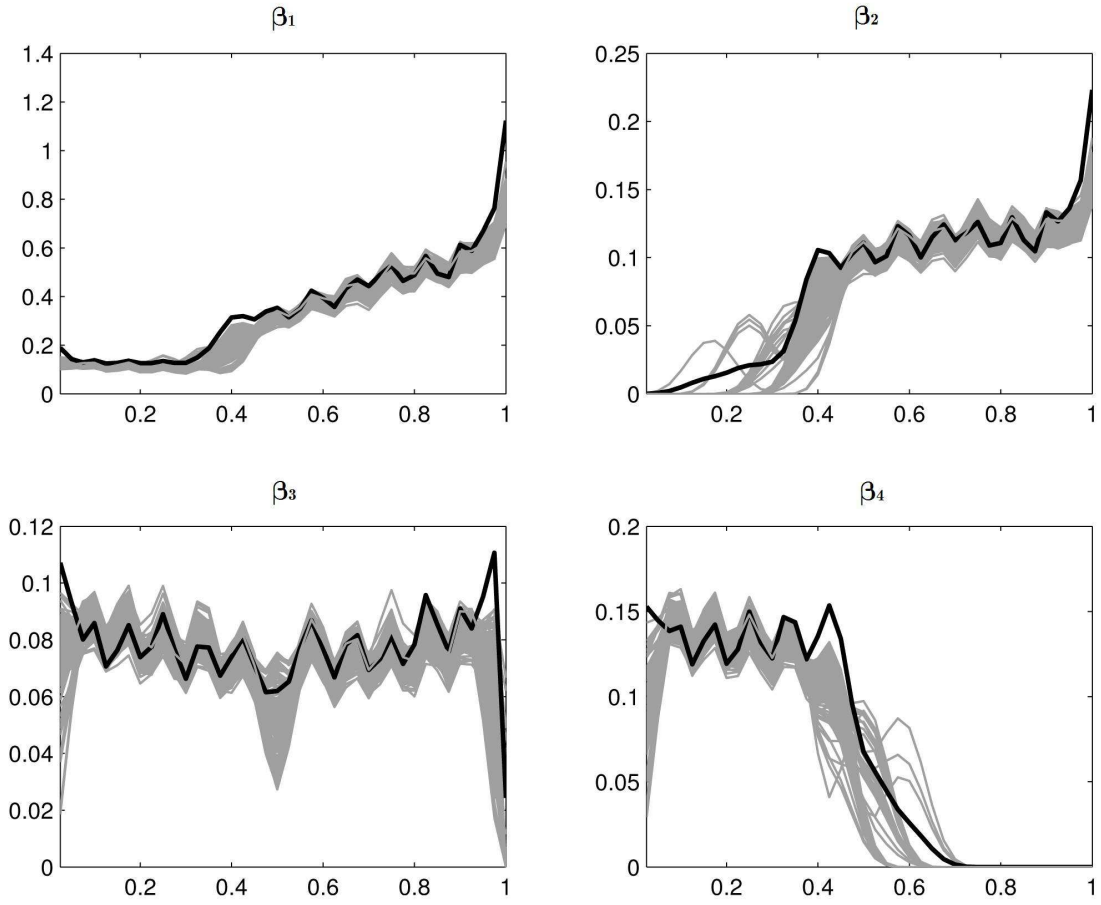

Figure 4.9: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions in Scenario 2.1.

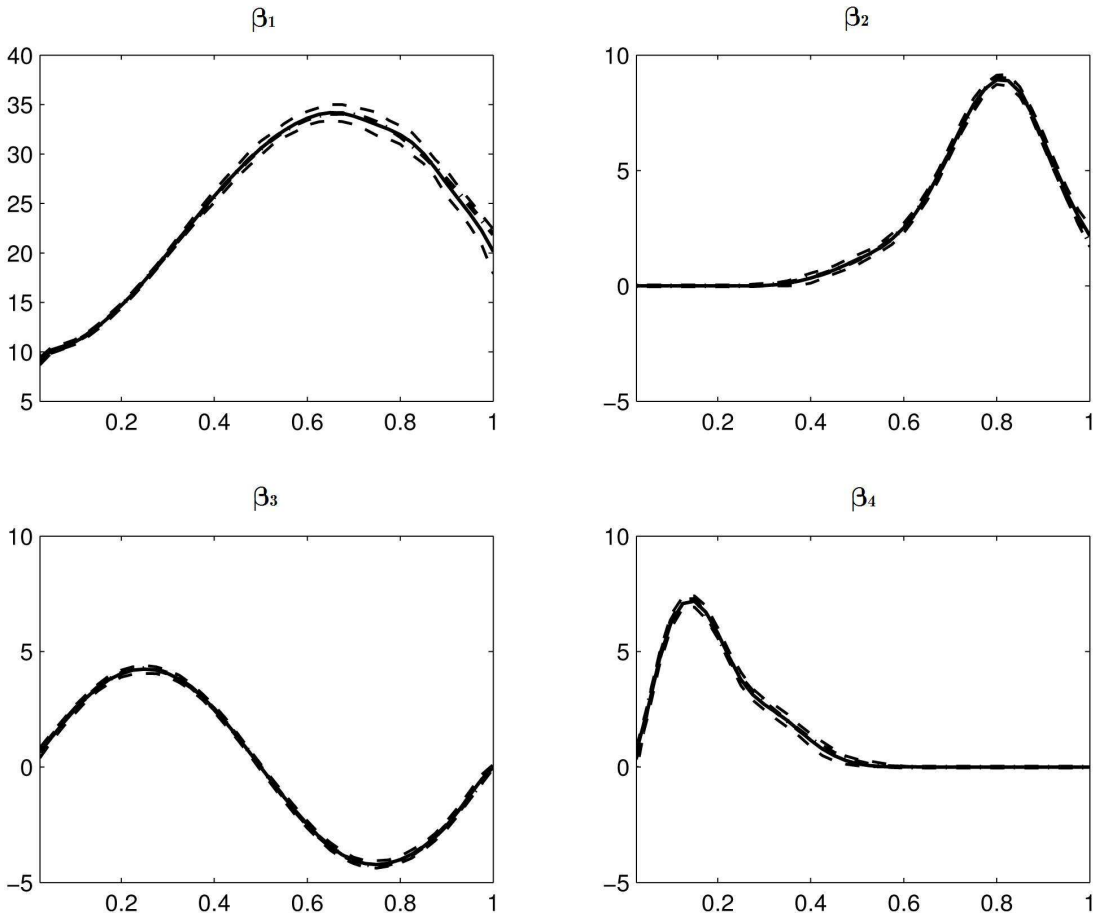

Figure 4.10: Empirical confidence intervals in Scenario 2.1.

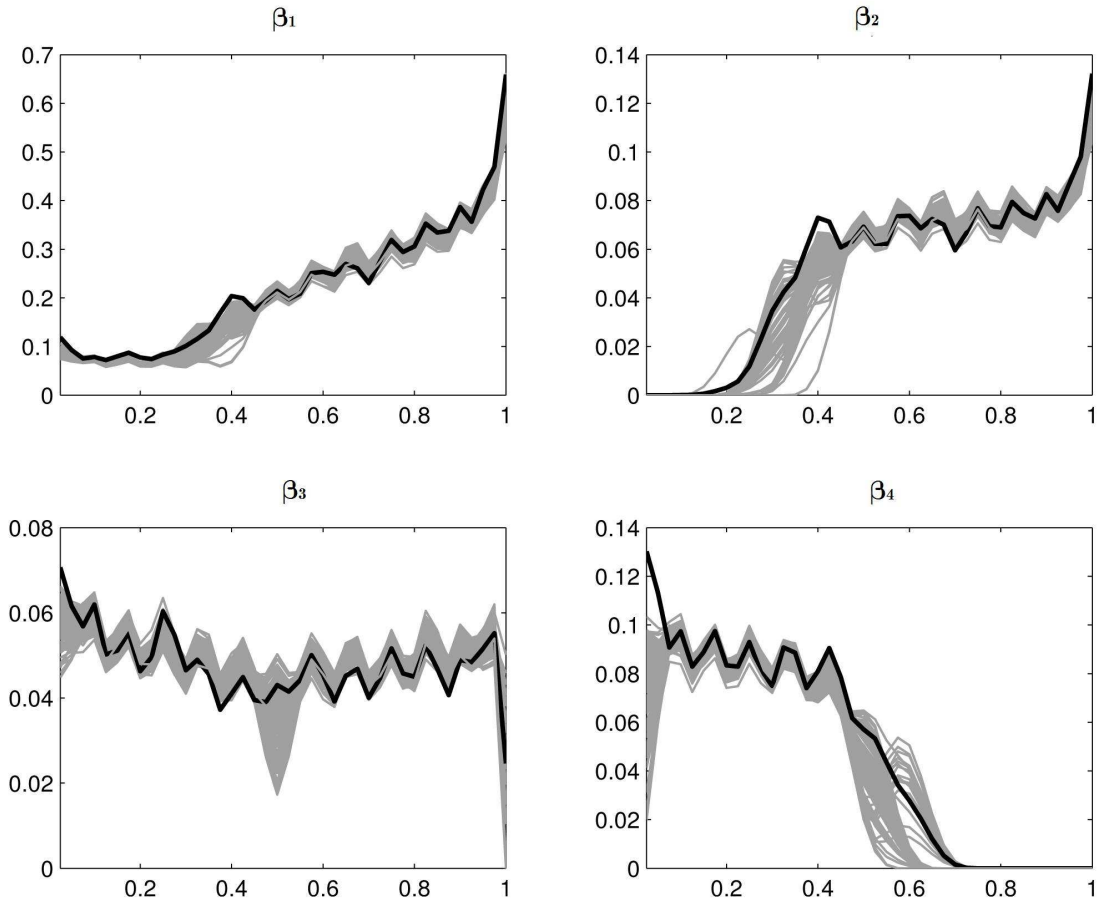

Figure 4.11: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions in Scenario 2.2.

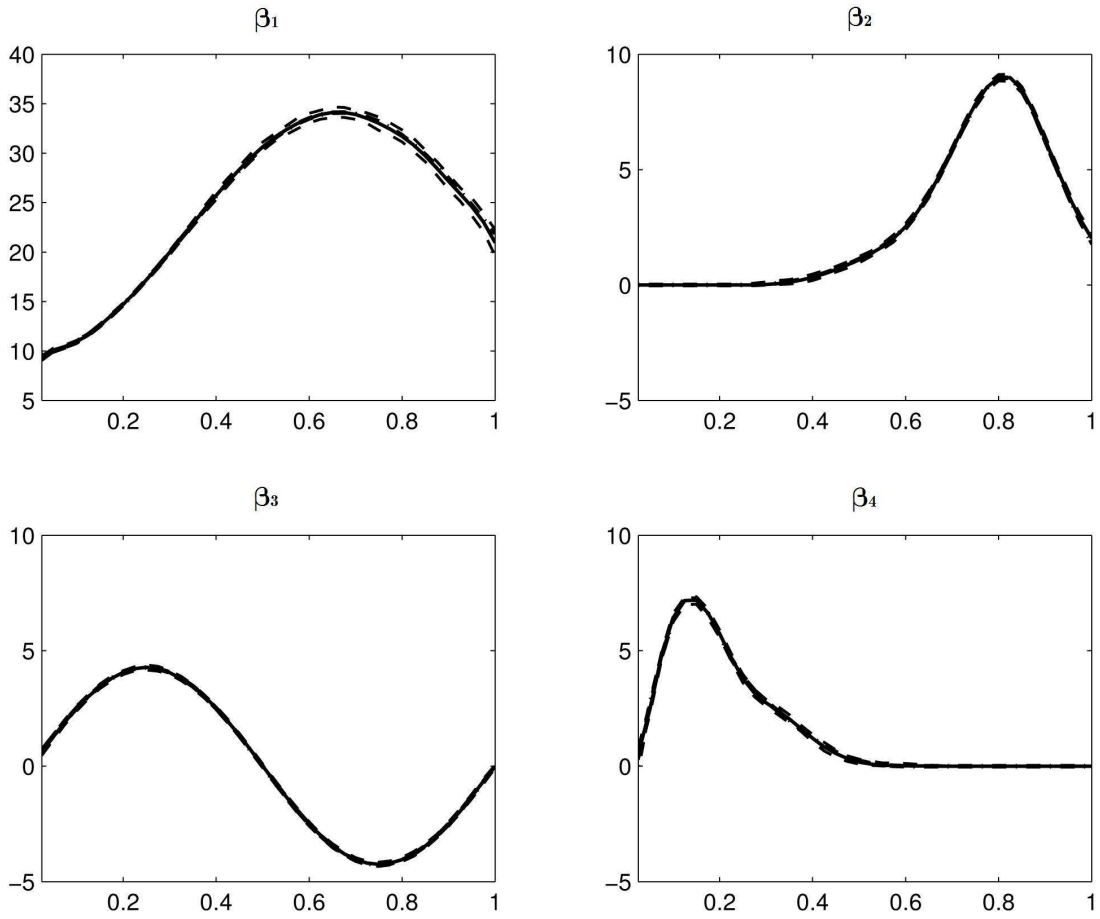

Figure 4.12: Empirical confidence intervals in Scenario 2.2.

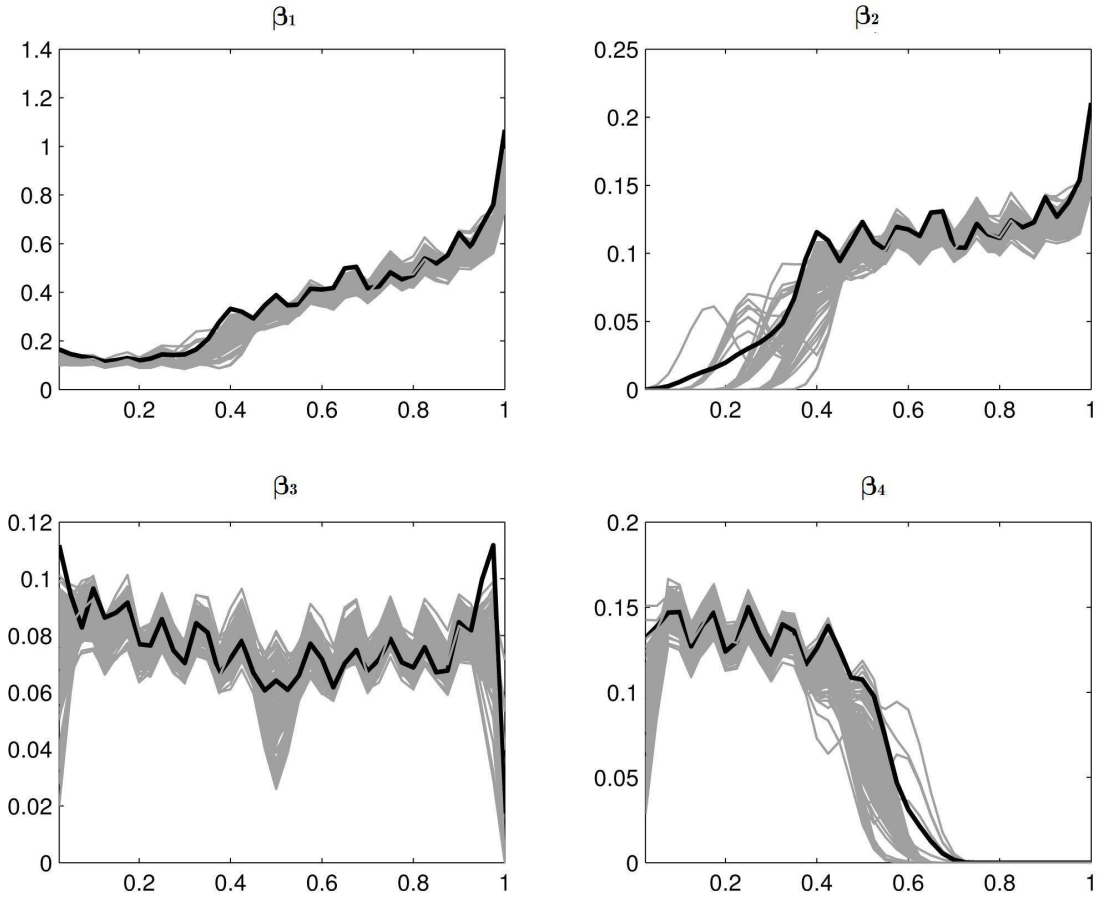

Figure 4.13: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions in Scenario 3.

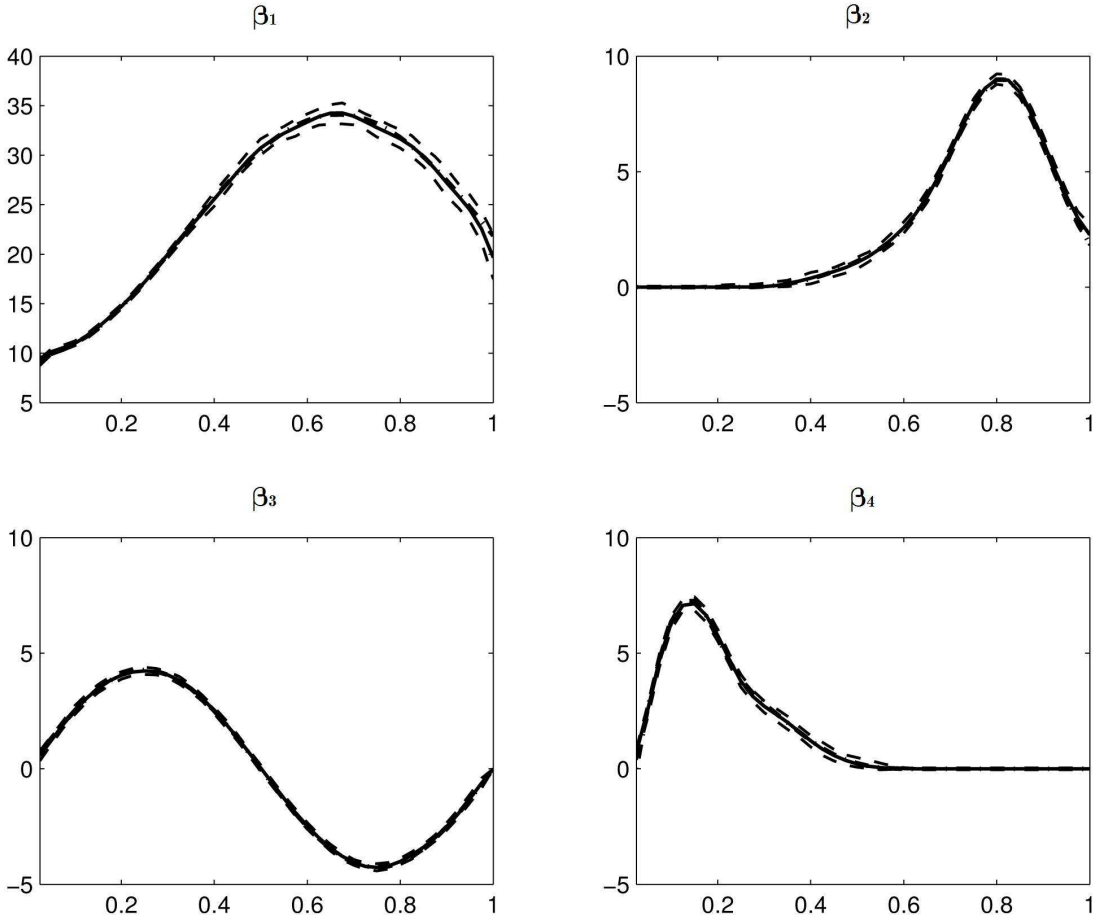

Figure 4.14: Empirical confidence intervals in Scenario 3.

CHAPTER 5

Real Data Analysis

We apply our proposed method to two real data sets, a gene expression data [Lee et al., 2002, Spellman et al., 1998] and the Boston housing data to illustrate the effectiveness of our method.

5.1 Application to Yeast Cell Cycle Gene Expression Data

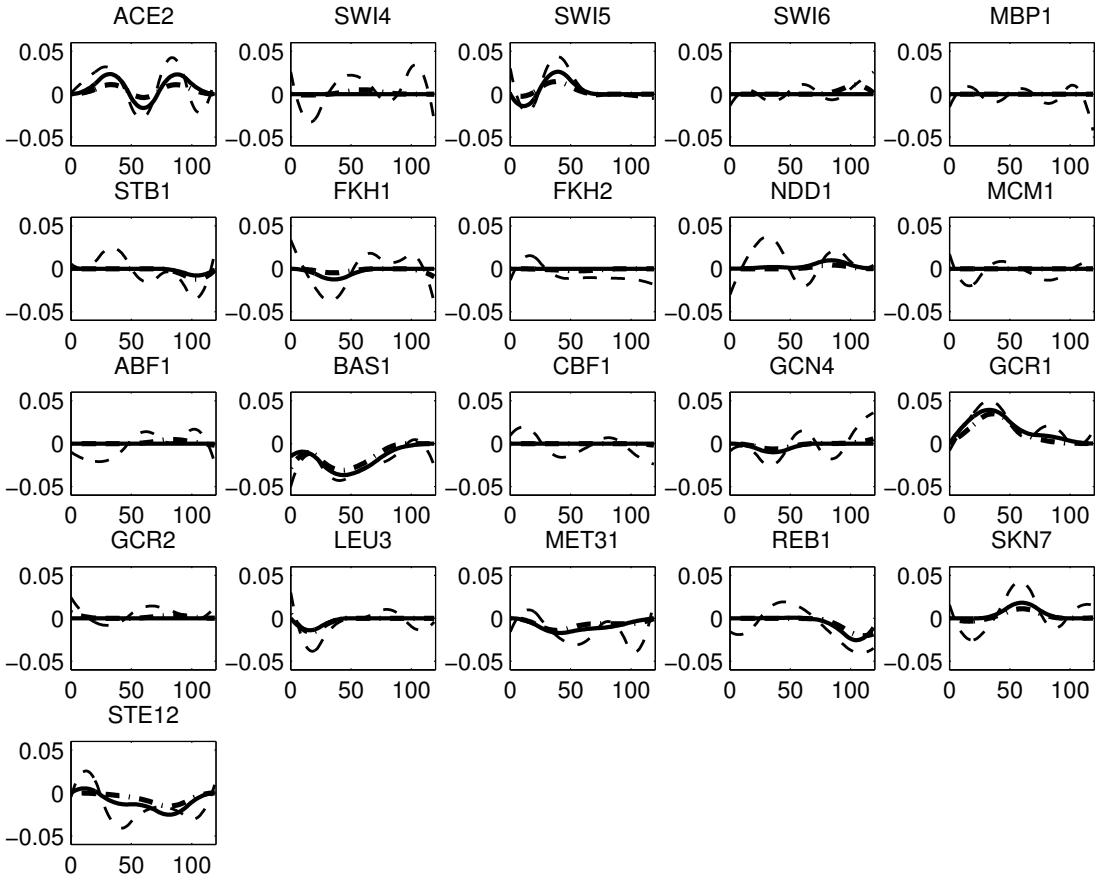

Figure 5.1: Subplots of estimated coefficient functions for the 21 confirmed TFs using LSE (dashed), Lasso (dot-dashed) and $\text{PLSE}_{0.5}$ (solid).

In biological sciences, gene expression data are frequently collected. Scientists believe that transcription factors (TFs) have effect on genome's cell cycle regulation. They have made

great effort in identifying key TFs in the regulatory network based on a set of gene expression measurements. In this study, we analyze the relationship between the level of gene expression and the physical binding of TFs from chromatin immunoprecipitation (ChIP-chip) data [Lee et al., 2002]. One of the gene expression data sets comes from an α factor synchronization experiment of 542 genes, in which mRNA levels are measured every 7 minutes during 119 minutes, resulting in 18 measurements in total [Spellman et al., 1998].

The ChIP-chip data contains the binding information of 106 transcription factors, among which 21 TFs are confirmed to be related to cell cycle regulation by experiment. Wang et al. [2007] demonstrated that a variable selection procedure is able to identify some of those key TFs. It is believed that the effects of TFs vary during the cell cycle. In Chun and Keleş [2010], the authors considered a varying coefficient model to study which TFs are important in gene expression. But they did not focus on the active periods of TFs, which is reflected in local sparsity of the coefficient functions. In this dissertation we apply our method to identify the key TFs and estimate the effects of those selected TFs over time. In addition, our approach allows us to investigate whether active and inactive periods during the cycle could be identified for each TF. Let y_{kt} denote the gene expression level for gene k at time t for $k = 1, \ldots, 542$ and $t = 1, \ldots, 18$, and let x_{ki} denote the binding information of transcription factor i for gene k, for $i = 1, \ldots, 106$. Then the varying coefficient model can be written as

$$y_{kt} = \beta_0(t) + \sum_{i=1}^{106} \beta_k(t) x_{ik} + \epsilon_{kt},$$

where $\beta_i(t)$ models the effect of the i-th transcription factor on gene expression at time t, and for the k-th gene ϵ_{kt}'s are independent over time. Similar to the simulation study, we apply our method together with LSE and Lasso methods and compared the identification of active period of each TF within the cell cycle process. Each coefficient function is approximated with quadratic B-splines defined on time interval $[0, 119]$ with four equally spaced knots. The number of knots is selected by cross-validation. For easy comparison, we standardize x_{ik}. It is not surprising that LSE does not identify the key TFs and selects all TFs. Lasso

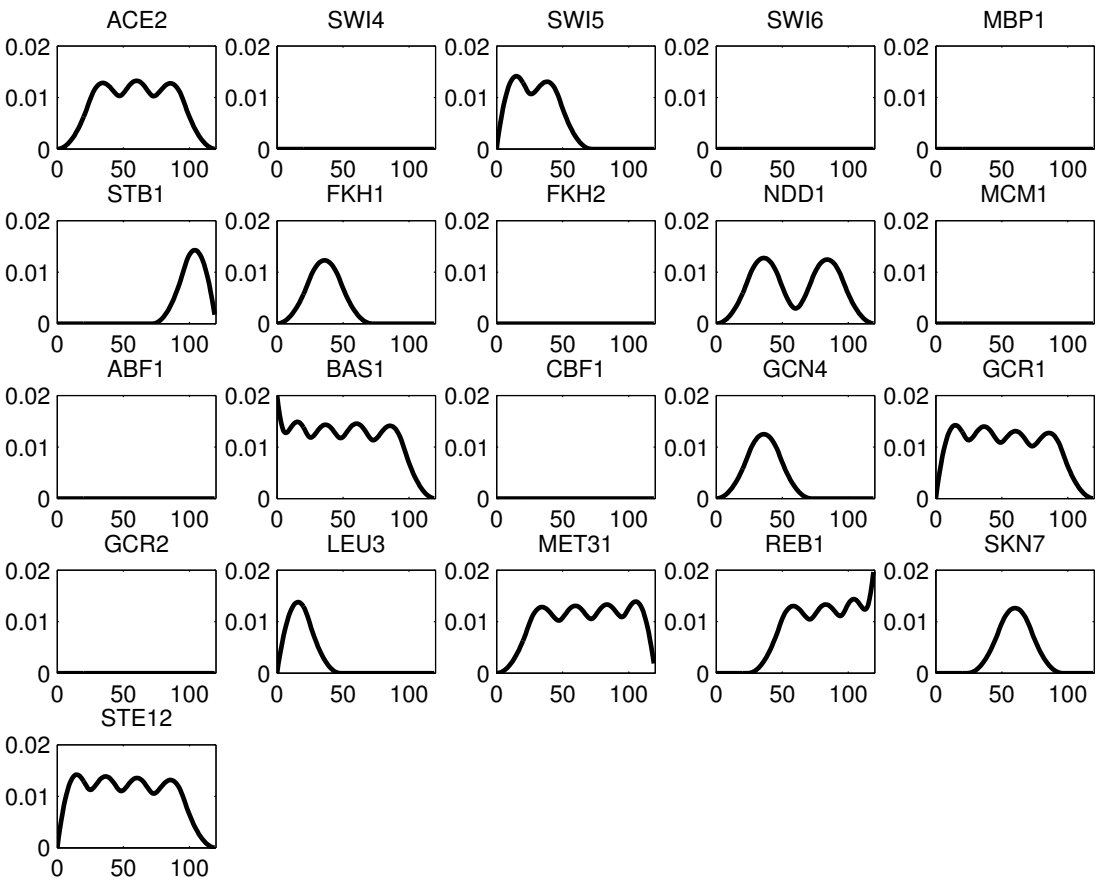

Figure 5.2: Asymptotic standard errors for confirmed TFs in yeast gene expression data.

method identifies 102 TFs as important, while our proposed method identifies 66 TFs. In Figure 5.1, the estimated coefficient functions for 21 experimentally confirmed TFs are shown, and corresponding standard error estimates are displayed in Figure 5.2. From this figure, we could tell 13 of them are selected by our proposed method while 20 of them are selected by Lasso method. In Chun and Keleş [2010], the authors selected 48 TFs, 15 of which are verified TFs. In addition, our proposed method identifies some inactive period for selected TFs. For example, SW15, FKH1, GCN4 and LEU3 are selected TFs which are inactive for the later 60 minutes, and STR1 and REB1 are selected TFs which have some latency in joining the expression. And for SKN7 is believed to be inactive at the beginning and at the end of the process.

5.2 Application to Boston Housing Data

Boston housing data is one of the widely used data sets in statistics literature. It is publicly available at `http://lib.stat.cmu.edu/datasets/boston`. This data set includes 506 median values of owner-occupied homes in Boston area and 12 potential variables that may impact housing value. Here we consider seven explanatory variables as suggested in earlier studies [e.g., Fan and Huang, 2005]. These variables are `CRIM` (per capita crime rate by town), `RM` (average number of rooms per dwelling), `TAX` (full-value property-tax rate per $\$10,000$), `NOX` (nitric oxides concentration in parts per 10 million), `PTRATIO` (pupil-teacher ratio by town), `AGE` (proportion of owner-occupied units built prior to 1940), and variable `B` ($1000(Bk - 0.63)^2$ where Bk is the proportion of blacks in town). Their influences on housing value are assumed to vary with the level of `LSTAT`, the percentage of lower status of the population. Similar studies with varying coefficient partially linear model were carried out in Fan and Huang [2005] and Leng [2009]. The authors studied function estimation and model selection, while their primary interest center on selecting the variables that are important over the entire range of `LSTAT`, i.e., identifying global sparsity. Here, we are interested in the question whether the impact of those variables on housing value when `LSTAT` is large is as significant as the impact when `LSTAT` is small. This could be referred to as local sparsity of the coefficient functions. In this study, such situation is taken into account while fitting the model. Let t be the scaled $\sqrt{\texttt{LSTAT}}$ on interval $[0, 1]$ as an indicator of `LSTAT`, and write the model as

$$y_{kt} = \sum_{i=1}^{7} \beta_i(t) x_{ki} + \epsilon_{kt},$$

where y_{kt} represents the k-th median value, x_{ki} for $i = 1, \ldots, 7$ are the standardized explanatory variables `CRIM`, `RM`, `NOX`, `PTRATIO`, `AGE`, and `B` separately, and $\beta_i(t)$ is the coefficient function of level t for the i-th variable. Figure 5.3 presents estimated coefficient functions by our procedure and by Lasso method, and Figure 5.4 presents the corresponding asymptotic

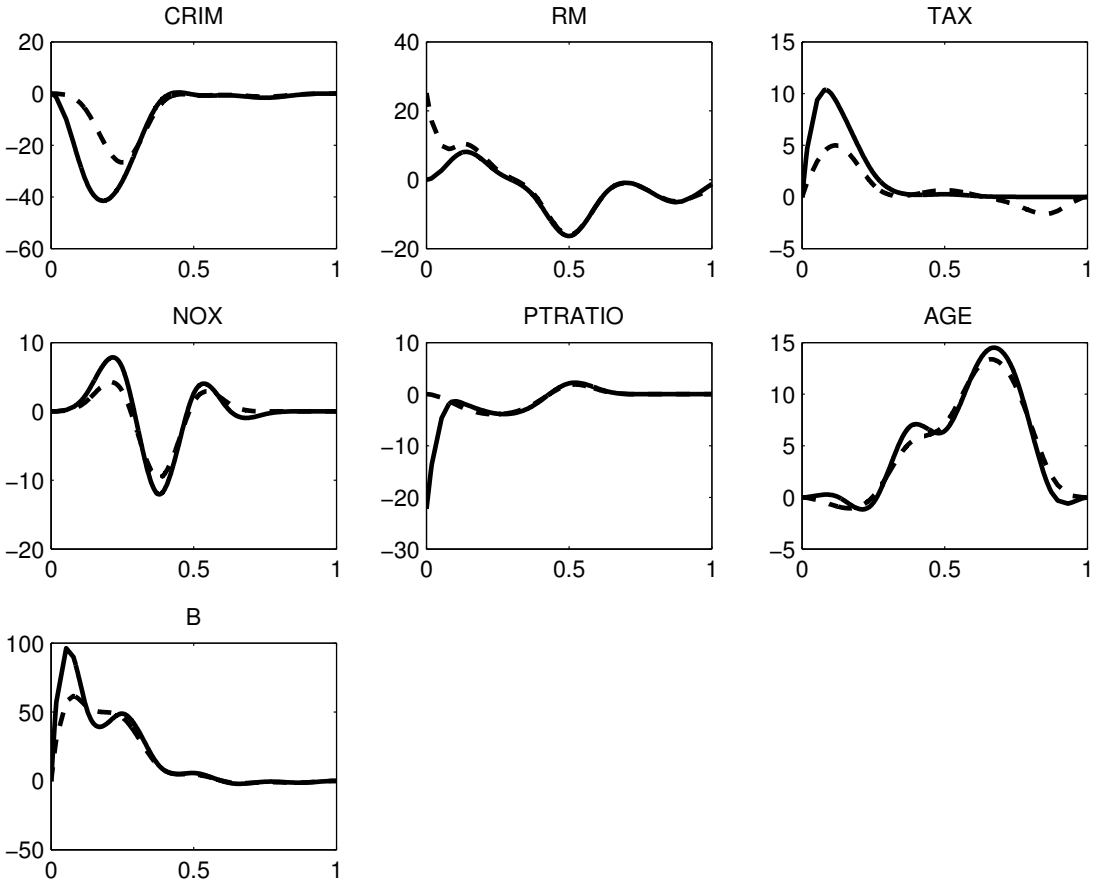

Figure 5.3: Subplots of estimated coefficient functions for the covariates in Boston housing data using Lasso (dot-dashed) and $\text{PLSE}_{0.5}$ (solid).

standard errors. In general, it shows that the crime rate (CRIM) has a negative effect on housing price, while the effect of tax rate (TAX), proportion of houses built prior to 1940 (AGE) and index of proportion of blacks (B) have positive effect. The effect of the average number of rooms per house (RM) and NOX is sensitive to LSTAT levels, and the effect of RM on housing price is negative most of time. Except for RM and AGE, the effect of the remaining variables varies a lot depending on the LSTAT levels among wealthy population, but is not significant among poor population. For the variable of PTRATIO, in contrast to the result of monotone increasing impact on housing price from Leng [2009], we believe its effect will reduce to zero when LSTAT is sufficiently large, i.e, among poor population. Similarly, the impact of TAX in Fan and Huang [2005] is modeled as monotone decreasing, but in our analysis it appears to have no effect on housing price when LSTAT is above certain level. This applies to the variable NOX as well.

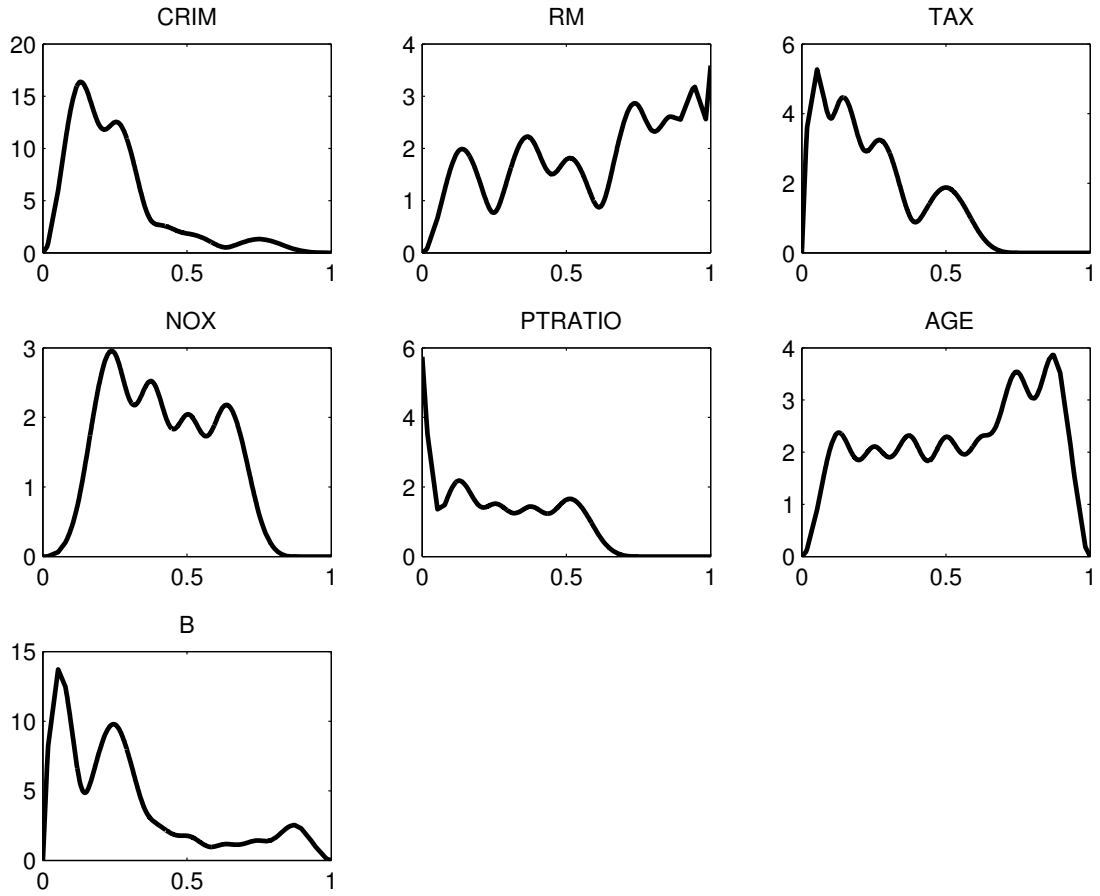

Figure 5.4: Asymptotic standard errors for Boston housing data.

CHAPTER 6

Technical assumptions and proofs

6.1 Technical assumptions

The following assumptions are made for Theorems 1, 2 and 3 in Chapter 3. These are standard assumptions used to establish asymptotic properties of nonparametric estimation procedures for varying coefficient models; see Huang et al. [2004] and Wang et al. [2008] for more details.

(A1) The response and covariate processes $\{y_k(t), \boldsymbol{x}_k(t),\, k = 1, \ldots, n\}$ are iid as $\{y(t), \boldsymbol{x}(t)\}$. And the observation time points, $t_{kl},\ l = 1, \ldots, n_k,\ k = 1, \ldots, n$, are iid from an unknown density, $f(t)$, on $[0, M]$, where $f(t)$ is uniformly bounded away from zero and infinity. That is, $0 < h_1 \leq f(t) \leq h_2 < \infty$ for some positive constants h_1 and h_2. Moreover, the observation time points are independent of the response and covariate processes $\{y_k(t), \boldsymbol{x}_k(t),\, k = 1, \ldots, n\}$.

(A2) The eigenvalues of the matrix $E[\boldsymbol{x}(t)\boldsymbol{x}^T(t)]$ are uniformly bounded away from zero and infinity for $t \in [0, M]$, that is, there exist positive constants M_1 and M_2 to be the lower and upper bound of the eigenvalues for all $t \in [0, M]$.

(A3) There exists a positive constant M_3 such that $|x_i(t)| \leq M_3$ for $t \in [0, M]$ and $i = 1, \ldots, p$.

(A4) There exists a positive constant M_4 such that $E\{\epsilon^2(t)\} \leq M_4$ for all $t \in [0, M]$

(A5) $\text{limsup}_n(\max_i K_i / \min_i K_i) < \infty$.

6.2 Proof of Theorem 1

The following lemma from Lemma A.3 of Huang et al. [2004] will be used in the proof.

Lemma 4. *Suppose that* $\lim_{n\to\infty} K_n \log K_n/n = 0$. *There are positive constants* C_1 *and* C_2 *such that, except on an event whose probability tends to zero, all eigenvalues of* $n^{-1}K_n\boldsymbol{U}^T\boldsymbol{U}$ *fall between* C_1 *and* C_2, *and consequently* $\boldsymbol{U}^T\boldsymbol{U}$ *is invertible.*

Proof of Theorem 1. Note that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le \|\widetilde{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_2 + \|\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}^0\|_2.$$

By B-spline property, $\|\beta_i - \widetilde{\beta}_i\|_2 = O_p(K_n^{-2})$ where $\widetilde{\beta}_i$ is an approximation in B-spline space as defined in (2.3). It can be shown that the same rate holds true if $\widetilde{\beta}_i$ is replaced by its sparse approximation of $\widetilde{\beta}_i^0$ (see Lemma 1 in Wang and Kai [2015]). Thus, $\|\widetilde{\boldsymbol{\beta}}^0 - \boldsymbol{\beta}\|_2 = O_p(K_n^{-2})$.

For the second term, by (A5) and B-spline property, we have $\|\widetilde{\beta}_i\|_2^2 \le D_i\|\boldsymbol{\alpha}_i\|_2^2/K_n$ for some positive constant D_i, $i = 1, \ldots, p$ [de Boor, 2001, Huang et al., 2004]. Denote $D_* = \max_i D_i$, and we have

$$\|\widetilde{\boldsymbol{\beta}}^0 - \widehat{\boldsymbol{\beta}}\|_2^2 = \sum_{i=1}^{p} \|\widetilde{\beta}_i^0 - \widehat{\beta}_i\|_2^2 \le \sum_{i=1}^{p} \frac{D_i}{K_n}\|\widetilde{\boldsymbol{\alpha}}_i^0 - \widehat{\boldsymbol{\alpha}}_i\|_2^2 \le D_* \frac{\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2}{K_n}$$

Therefore,

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 = O_p\Big(K_n^{-4} + \frac{\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2}{K_n}\Big).$$

Below we concentrate on the term $\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2$ and in particular we show that $\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2 = O_p(n^{-1}K_n^2)$.

By the minimality of $\widehat{\boldsymbol{\alpha}}$, we have $\mathrm{pl}(\widehat{\boldsymbol{\alpha}}) \le \mathrm{pl}(\widetilde{\boldsymbol{\alpha}}^0)$; that is,

$$\|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2 - \|\boldsymbol{y} - \boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 \le \lambda_n \sum_{i=1}^{p}\sum_{g=1}^{G_i} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma - \lambda_n \sum_{i=1}^{p}\sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma. \tag{6.1}$$

Note that, the right hand side of (6.1) can be decomposed into two terms,
$\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma - \lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma$ and $\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma -$
$\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma$. For the first term, applying the inequality $|b^\gamma - a^\gamma| \le 2|b-a|b^{\gamma-1}$, for $a, b \ge 0$, and Cauchy-Schwarz inequality yields that

$$
\begin{aligned}
&\left|\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right| \\
\leq\ & 2\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\left|\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1-\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1\right|\cdot\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma-1} \\
\leq\ & 2\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1\cdot\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma-1} \\
\leq\ & 2(d+1)^{1/2}\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma-1}\cdot\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2 \\
\leq\ & 2(d+1)^{1/2}\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{2(\gamma-1)}\right)^{1/2}\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2^2\right)^{1/2}.
\end{aligned}
$$

For the second term, note that $\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1=0$ for $g\in\mathcal{A}_{i2}$. Thus, the second term is less than or equal to zero. Combining above results and (6.1), we have

$$
\begin{aligned}
&\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 \\
&\leq\lambda_n\left|\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right|+\lambda_n\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2^{\gamma}\right) \\
&\leq\lambda_n\left|\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right| \\
&\leq 2\lambda_n\phi_n\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2^2\right)^{1/2}
\end{aligned}
$$

It follows that

$$
\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2\leq 2\lambda_n\phi_n(d+1)^{1/2}\|\widetilde{\boldsymbol{\alpha}}^0-\widehat{\boldsymbol{\alpha}}\|_2, \tag{6.2}
$$

where $\phi_n=(d+1)^{1/2}\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{2(\gamma-1)}\right)^{1/2}$.

On the other hand, straightforward calculation gives that

$$
\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 &= (\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}\widehat{\boldsymbol{\alpha}}-(\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0)^T\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0-2\boldsymbol{y}^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0) \\
&= (\boldsymbol{U}\widehat{\boldsymbol{\alpha}}+\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0-2\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0-2\boldsymbol{\epsilon}_*)^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0) \\
&= \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\|_2^2-2\boldsymbol{\epsilon}_*^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0) \\
&\geq \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\|_2^2-2\left|\boldsymbol{\epsilon}_*^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\right|,
\end{aligned}
$$

where $\boldsymbol{\epsilon}_* = \boldsymbol{\epsilon} - \boldsymbol{e}$, $\boldsymbol{e} = (\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_n^T)^T$ with

$$\boldsymbol{e}_k = (\mathcal{U}_k(t_{k1})\widetilde{\boldsymbol{\alpha}}^0 - \boldsymbol{x}_k(t_{k1})^T\boldsymbol{\beta}(t_{k1}), \ldots, \mathcal{U}_k(t_{kn_k})\widetilde{\boldsymbol{\alpha}}^0 - \boldsymbol{x}_k(t_{kn_k})^T\boldsymbol{\beta}(t_{kn_k}))^T .$$

Let $\delta_n = \|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2$, then by Lemma 4, $\|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0)\|_2^2 \geq C_1 n K_n^{-1}\delta_n^2$ with probability approaching 1. In addition, applying Cauchy-Schwarz inequality yields that $(\boldsymbol{\epsilon}_*^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0))^2 \leq \delta_n^2(\boldsymbol{\epsilon}_*^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}_*)$. Further, $E(\boldsymbol{\epsilon}_*^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}_*) = E(\boldsymbol{\epsilon}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}) + E(\boldsymbol{e}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{e})$. As a consequence of Lemma A.3 of Wang et al. [2008], we have $E(\boldsymbol{\epsilon}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}) = O(n)$ with n_k uniformly bounded. Similarly, we have $E(\boldsymbol{e}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{e}) = O(n)$ since $E(e(t_{kl})e(t_{kl'})) \leq C\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}^0\|_\infty^2$ for some constant C and $\|\boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}}^0\|_\infty$ is bounded by $O(K_n^{-2})$. Therefore, $E(\boldsymbol{\epsilon}_*^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}_*) = O(n)$. Thus, we have

$$\|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2 - \|\boldsymbol{y} - \boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 \geq C_1 n {K_n}^{-1}\delta_n^2 - \delta_n O_p(n^{1/2}). \tag{6.3}$$

Combining (6.2) and (6.3), we have

$$\frac{nC_1}{K_n}\delta_n^2 - \delta_n O_p(n^{1/2}) \leq 2\lambda_n\phi_n(d+1)^{1/2}\delta_n,$$

and by (S1) we have $\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2 = O_p(n^{-1}K_n^2)$.

□

6.3 Proof of Theorem 2

Proof. First, for any i, define $\widehat{\alpha}_{ij}^*$ in the following way. Let $\widehat{\alpha}_{ij}^* = 0$ if $\{j-d, \ldots, j\} \cap \mathcal{A}_{i2} \neq \emptyset$, otherwise, $\widehat{\alpha}_{ij}^* = \widehat{\alpha}_{ij}$. Note that $\widehat{\boldsymbol{\alpha}}_{A_{ig}}^* = \boldsymbol{0}$ for $g \in \mathcal{A}_{i2}$.

By Karush-Kuhn-Tucker conditions, for $\widehat{\alpha}_{ij} \neq 0$ we have

$$2(\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T U_{(ij)} = \sum_{g=j-d}^{j} \gamma\lambda_n\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}\mathrm{sgn}(\widehat{\alpha}_{ij}),$$

where $U_{(ij)}$ is the column of $\boldsymbol{U}$ corresponding to $\widehat{\alpha}_{ij}$. Multiplying both sides by $(\widehat{\alpha}_{ij} - \widehat{\alpha}_{ij}^*)$

yields

$$\begin{aligned}
2(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*) &= \sum_{i,j}\sum_{g=j-d}^{j}\gamma\lambda_n\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}\mathrm{sgn}(\widehat{\alpha}_{ij})(\widehat{\alpha}_{ij}-\widehat{\alpha}^*_{ij}) \\
&= \gamma\lambda_n\sum_{i,j}\sum_{g\in\mathcal{A}_{i2}\cap\{j-d,\ldots,j\}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}|\widehat{\alpha}_{ij}|\cdot \\
&= \gamma\lambda_n\sum_{i=1}^{p}\sum_{g=1}^{G_i}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}(\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1-\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1).
\end{aligned}$$

Note that, $(\widehat{\alpha}_{ij}-\widehat{\alpha}^*_{ij})\mathrm{sgn}(\widehat{\alpha}_{ij})=|\widehat{\alpha}_{ij}|$ if $\{j-d,\ldots,j\}\cap\mathcal{A}_{i2}\neq\emptyset$.

Since $\gamma b^{\gamma-1}(b-a)\le b^\gamma-a^\gamma$ for $0\le a\le b$, we have, for $g\in\mathcal{A}_{i1}$,

$$\gamma\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}(\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1-\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1)\le\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}-\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma}.$$

Consequently, we have

$$2\left|(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*)\right|\le\lambda_n\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}(\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}-\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma})+\gamma\lambda_n\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}. \quad (6.4)$$

By the minimality of $\widehat{\boldsymbol{\alpha}}$, we have

$$\lambda_n\sum_{i=1}^{p}\sum_{g=1}^{G_i}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}-\lambda_n\sum_{i=1}^{p}\sum_{g=1}^{G_i}\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma}\le\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}^*\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2.$$

Note that $\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1=0$ for $g\in\mathcal{A}_{i2}$. Thus, we have $\sum_{i=1}^{p}\sum_{g=1}^{G_i}\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma}=\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma}$, and

$$\begin{aligned}
&2\left|(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*)\right|+(1-\gamma)\lambda_n\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma} \\
&\le\lambda_n\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}(\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}-\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma})+\lambda_n\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma} \\
&=\lambda_n\sum_{i=1}^{p}\sum_{g=1}^{G_i}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}-\lambda_n\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^{\gamma} \\
&\le\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}^*\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2 \\
&=\|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}^*-\widehat{\boldsymbol{\alpha}})\|_2^2+2(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*).
\end{aligned}$$

By Lemma 4 we have

$$\begin{aligned}
&(1-\gamma)\lambda_n \sum_{i=1}^{p} \sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma} \\
&\leq \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}})\|_2^2 + 2(\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*) - 2\left|(\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*)\right| \\
&\leq \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}})\|_2^2 \\
&\leq \frac{nC_2}{K_n}\|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2^2
\end{aligned}$$

Note that $\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}} = \mathbf{0}$ for $g \in \mathcal{A}_{i2}$. Thus, we have $\|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2^2 \leq \|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2$, and

$$(1-\gamma)\lambda_n \sum_{i=1}^{p} \sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma} \leq \frac{nC_2}{K_n}\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2 = O_p(K_n)$$

Since

$$\sum_{i=1}^{p} \sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma} \geq \left(\sum_{i=1}^{p} \sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1\right)^{\gamma} \geq \|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2^{\gamma},$$

then if $\|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2 > 0$, we have

$$\begin{aligned}
(1-\gamma)\lambda_n &\leq \frac{nC_2}{K_n}\|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2^2 \left\{\sum_{i=1}^{p} \sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right\}^{-1} \\
&\leq \frac{nC_2}{K_n}\|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2^{2-\gamma} \\
&\leq O_p(n^{\gamma/2}K_n^{1-\gamma}),
\end{aligned}$$

and thus

$$\Pr\left\{\|\widehat{\boldsymbol{\alpha}}^* - \widehat{\boldsymbol{\alpha}}\|_2^2 > 0\right\} \leq \Pr\left\{\frac{\lambda_n}{n^{\gamma/2}K_n^{1-\gamma}} \leq O_p(1)\right\}.$$

By assumption (S2), the right hand side converges to zero as n go to infinity, which implies that $(\widehat{\boldsymbol{\alpha}}_{A_{ig}} : g \in \mathcal{A}_{i2}) = \mathbf{0}$ with probability approaching to 1. □

6.4 Proof of Theorem 3

We will generalize Lemma 4 above to Lemma 4′ for the case of diverging p.

Lemma 4′. *Suppose that* $\lim_n p^2 K_n(\log p + \log K_n)/n = 0$. *There are positive constants* C_1 *and* C_2 *such that, except on an event whose probability tends to zero, all eigenvalues of* $n^{-1}K_n\boldsymbol{U}^T\boldsymbol{U}$ *fall between* C_1 *and* C_2.

We will begin with the introduction of some new notations.

(i) For sequence of positive numbers a_n and b_n, we call $a_n \asymp b_n$ if both a_n/b_n and b_n/a_n are bounded.

(ii) For $\boldsymbol{\beta}^{(1)}(t) = (\beta_1^{(1)}(t), \ldots, \beta_p^{(1)}(t))^T$ and $\boldsymbol{\beta}^{(2)}(t) = (\beta_1^{(2)}(t), \ldots, \beta_p^{(2)}(t))^T$, we define empirical inner product as

$$\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle_n = \frac{1}{n} \sum_k \frac{1}{n_k} \sum_l \left(p^{-1/2} \sum_i x_i(t_{kl}) \beta_i^{(1)}(t_{kl}) \right) \left(p^{-1/2} \sum_i x_i(t_{kl}) \beta_i^{(2)}(t_{kl}) \right),$$

and the theoretical inner product as

$$\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle = E \left[\left(p^{-1/2} \sum_i x_i(t) \beta_i^{(1)}(t) \right) \left(p^{-1/2} \sum_i x_i(t) \beta_i^{(2)}(t) \right) \right].$$

Denote the corresponding norms as $\|\boldsymbol{\beta}\|_n^2 = \langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle_n$ and $\|\boldsymbol{\beta}\|^2 = \langle \boldsymbol{\beta}, \boldsymbol{\beta} \rangle$.

In Huang et al. [2004], the authors considered similar notions of empirical inner product and theoretical inner product for finite p. Inspired by their work, we rescale both quantities by $1/p$. As will be seen later, such treatment is essential for the new theoretical development. Then we generalize the results of Lemma A.1 and Lemma A.2 in Huang et al. [2004] to Lemma 5 and Lemma 6 respectively as below.

Lemma 5. *Let* $\beta_i(t) = \sum_j \alpha_{ij} B_{ij}(t)$ *and* $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$, *then* $\|\boldsymbol{\beta}\|^2 \asymp \sum_{i=1}^p \|\beta_i\|_2^2/p \asymp \|\boldsymbol{\alpha}\|_2^2/(pK_n)$.

Proof of Lemma 5. Note that

$$\begin{aligned} \|\boldsymbol{\beta}\|^2 &= E \left[\frac{1}{p} \left(\sum_i x_i(t) \beta_i(t) \right)^2 \right] \\ &= \frac{1}{p} \int E \left(\sum_i x_i(t) \beta_i(t) \right)^2 f(t) dt \\ &= \frac{1}{p} \int \boldsymbol{\beta}^T(t) E \left(\boldsymbol{x}(t) \boldsymbol{x}^T(t) \right) \boldsymbol{\beta}(t) f(t) dt. \end{aligned}$$

Then by (A1) and (A2), we have $\|\boldsymbol{\beta}\|^2 \asymp 1/p \int \boldsymbol{\beta}^T(t) \boldsymbol{\beta}(t) dt \asymp 1/p \sum_{i=1}^p \|\beta_i\|_2^2$. By B-spline properties, $\|\beta_i\|_2^2 \asymp \|\boldsymbol{\alpha}_i\|_2^2/K_i$. Therefore, $\|\boldsymbol{\beta}\|^2 \asymp \|\boldsymbol{\alpha}\|_2^2/(pK_n)$ under (A5). □

Lemma 6. *Let* $\beta_i(t) = \sum_j \alpha_{ij} B_{ij}(t)$ *for* $i = 1, \ldots, p$, $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$, *and* $\mathbb{G}$ *be the collection of* $\boldsymbol{\beta}$. *Then*

$$P\left(\sup_{\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \in \mathbb{G}} \frac{|\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle_n - \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \rangle|}{\|\boldsymbol{\beta}^{(1)}\| \|\boldsymbol{\beta}^{(2)}\|} > s \right) \leq C p^2 K_n^2 \exp\left\{ \frac{-n}{p^2 K_n} \cdot \frac{s^2}{b_2 + b_3 s/p} \right\}, \quad s > 0$$

for some positive constant C, b_2 *and* b_3. *Further, if* $\lim_n p^2 K_n (\log p + \log K_n)/n = 0$, *then* $\sup_{\boldsymbol{\beta} \in \mathbb{G}} |\|\boldsymbol{\beta}\|_n^2 / \|\boldsymbol{\beta}\|^2 - 1| = o_{\mathrm{p}}(1)$.

Proof of Lemma 6. Let $\boldsymbol{B}_{ij}(t)$ be the p-dimensional vector with the i-th entry being $B_{ij}(t)$ and all other entries zero. Then

$$\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'} \rangle_n = \frac{1}{np} \sum_k \frac{1}{n_k} \sum_l [x_i(t_{kl}) B_{ij}(t_{kl}) x_{i'}(t_{kl}) B_{i'j'}(t_{kl})],$$

$$\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'} \rangle = \frac{1}{p} E[x_i(t) B_{ij}(t) x_{i'}(t) B_{i'j'}(t)].$$

Denote $R_{iji'j'}(t) = x_i(t) B_{ij}(t) x_{i'}(t) B_{i'j'}(t)$, then

$$\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'} \rangle_n - \langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'} \rangle = \frac{1}{np} \sum_k \frac{1}{n_k} \sum_l \left(R_{iji'j'}(t_{kl}) - E[R_{iji'j'}(t)] \right).$$

By the fact that $0 \leq B_{ij}(t) \leq 1$ for $t \in [0, M]$ and (A3), we have

$$|R_{iji'j'}(t_{kl}) - E[R_{iji'j'}(t)]| \leq |x_i(t_{kl})||x_{i'}(t_{kl})| + E[|x_i(t)||x_{i'}(t)|] \leq 2M_3^2.$$

By B-spline property, (A1) and (A5), we have

$$Var(R_{iji'j'}(t_{kl})) \leq \quad E(x_i^2(t) x_{i'}^2(t) B_{ij}^2(t) B_{i'j'}^2(t)) \leq \quad C h_2 M_3^4 K_n^{-1},$$

for some positive constant C. Then, as a consequence of Bernstein's inequality, we have, for $s > 0$,

$$P\left(|\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'} \rangle_n - \langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'} \rangle| > s \right) \leq b_1 \exp\left\{ -\frac{(nps)^2}{b_2 n K_n^{-1} + b_3 nps} \right\},$$

for some positive constants b_1, b_2 and b_3.

Let Ω_n be the event on which $|\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'}\rangle_n - \langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'}\rangle| \le s/(p^2 K_n)$ for all $j = 1, \ldots, (K_i + d + 1)$, $j' = 1, \ldots, (K_{i'} + d + 1)$, and $i, i' = 1, \ldots, p$. Then

$$P\left(\Omega_n^c\right) \le Cp^2 K_n^2 \exp\left\{-\frac{n}{p^2 K_n} \cdot \frac{s^2}{b_2 + b_3 s/p}\right\},$$

for some constant C. For $k = 1, \ldots, n$, the n_k are uniformly bounded.

For $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)} \in \mathbb{G}$, we have

$$|\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle_n - \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle| = \left|\sum_{i,j}\sum_{i',j'} \alpha_{ij}^{(1)} \alpha_{i'j'}^{(2)} (\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'}\rangle_n - \langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'}\rangle)\right|.$$

Let $(i', j') \in A(i, j)$ if the intersection of the supports of $B_{i'j'}$ and B_{ij} contains an open interval. Then $\langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'}\rangle_n = \langle \boldsymbol{B}_{ij}, \boldsymbol{B}_{i'j'}\rangle = 0$ if $(i', j') \notin A(i, j)$. The cardinality of $A(i, j)$ is bounded by Cp for some constant C for all i, j. Then, on Ω_n, we have

$$\begin{aligned}
& |\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle_n - \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle| \\
\le\ & \sum_{i,j}\sum_{i',j'} |\alpha_{ij}^{(1)}||\alpha_{i'j'}^{(2)}| \frac{s}{p^2 K_n} I_{\{(i,j)\in A(i',j')\}} \\
\le\ & \frac{s}{p^2 K_n} \sum_{i,j} |\alpha_{ij}^{(1)}| \left(\sum_{i',j'} (|\alpha_{i'j'}^{(2)}| I_{\{(i,j)\in A(i',j')\}})^2\right)^{1/2} (Cp)^{1/2} \\
\le\ & \frac{s}{p^2 K_n} \left(\sum_{i,j} |\alpha_{ij}^{(1)}|^2\right)^{1/2} \left(\sum_{i,j}\sum_{i',j'} |\alpha_{i'j'}^{(2)}|^2 I_{\{(i,j)\in A(i',j')\}}\right)^{1/2} (Cp)^{1/2} \\
\le\ & \frac{Cs}{pK_n} \|\boldsymbol{\alpha}^{(1)}\|_2 \|\boldsymbol{\alpha}^{(2)}\|_2
\end{aligned}$$

It follows from Lemma 5, $|\langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle_n - \langle \boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}\rangle| \le Cs\|\boldsymbol{\beta}^{(1)}\|\|\boldsymbol{\beta}^{(2)}\|$ on Ω_n for some positive constant C. Therefore, the conclusion follows. □

Proof of Lemma 4′. By Lemma 5 and Lemma 6, we have $\|\boldsymbol{\alpha}\|_2^2/(pK_n) \asymp \|\boldsymbol{\beta}\|^2 \asymp \|\boldsymbol{\beta}\|_n^2 \asymp \boldsymbol{\alpha}^T \boldsymbol{U}^T \boldsymbol{U} \boldsymbol{\alpha}/(np)$ except on an event whose probability tends to zero. Then the desired result follows. □

Now we will prove Theorem 3 of consistency followed by sparsistency.

Proof of consistency in Theorem 3. Note that, for each i, $i = 1, \ldots, p$, we have

$$\|\widehat{\beta}_i - \beta_i\|_2 \leq \|\widetilde{\beta}_i^0 - \beta_i\|_2 + \|\widehat{\beta}_i - \widetilde{\beta}_i^0\|_2.$$

By B-spline property, $\|\beta_i - \widetilde{\beta}_i\|_2 = O_p(K_n^{-2})$ where $\widetilde{\beta}_i$ is an approximation in B-spline space as defined in (2.3). It can be shown that the same rate holds true if $\widetilde{\beta}_i$ is replaced by its sparse approximation $\widetilde{\beta}_i^0$ (see Lemma 1 in Wang and Kai [2015]). Thus, $\|\widetilde{\beta}_i^0 - \beta_i\|_2 = O_p(K_n^{-2})$. It is worth mentioning that, if $\beta_i = 0$, we can set $\widetilde{\beta}_i = 0$ and thus, $\widetilde{\beta}_i^0 = 0$.

For the second term, by (A5) and B-spline property, there exists some positive constant D for each $i = 1, \ldots, p$ [de Boor, 2001, Huang et al., 2004] such that

$$\|\widetilde{\beta}_i^0 - \widehat{\beta}_i\|_2^2 \leq \frac{D}{K_n} \|\widetilde{\boldsymbol{\alpha}}_i^0 - \widehat{\boldsymbol{\alpha}}_i\|_2^2.$$

Therefore,

$$\|\widehat{\beta}_i - \beta_i\|_2^2 = O_p\Big(K_n^{-4} + \frac{\|\widehat{\boldsymbol{\alpha}}_i - \widetilde{\boldsymbol{\alpha}}_i^0\|_2^2}{K_n}\Big),$$

and

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 = O_p\Big(pK_n^{-4} + \frac{\sum_{i=1}^p \|\widehat{\boldsymbol{\alpha}}_i - \widetilde{\boldsymbol{\alpha}}_i^0\|_2^2}{K_n}\Big),$$

Below we concentrate on the term $\sum_{i=1}^p \|\widehat{\boldsymbol{\alpha}}_i - \widetilde{\boldsymbol{\alpha}}_i^0\|_2 = \|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2$. In particular, we will show that $\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2 = O_p(n^{-1}K_n^2 p)$.

By the minimality of $\widehat{\boldsymbol{\alpha}}$, we have $\mathrm{pl}(\widehat{\boldsymbol{\alpha}}) \leq \mathrm{pl}(\widetilde{\boldsymbol{\alpha}}^0)$; that is,

$$\|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2 - \|\boldsymbol{y} - \boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 \leq \lambda_n \sum_{i=1}^p \sum_{g=1}^{G_i} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma - \lambda_n \sum_{i=1}^p \sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma. \tag{6.5}$$

Note that, the right hand side of (6.5) can be decomposed into two terms,
$\lambda_n \sum_{i=1}^p \sum_{g\in\mathcal{A}_{i1}} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma - \lambda_n \sum_{i=1}^p \sum_{g\in\mathcal{A}_{i1}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma$ and $\lambda_n \sum_{i=1}^p \sum_{g\in\mathcal{A}_{i2}} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma -$
$\lambda_n \sum_{i=1}^p \sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma$. For the first term, applying the inequality $|b^\gamma - a^\gamma| \leq 2|b-a|b^{\gamma-1}$, for $a, b \geq 0$, and Cauchy-Schwarz inequality yields that

$$\left| \sum_{i=1}^p \sum_{g\in\mathcal{A}_{i1}} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^\gamma - \sum_{i=1}^p \sum_{g\in\mathcal{A}_{i1}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma \right|$$

$$
\begin{aligned}
&\leq 2\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\left|\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1-\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1\right|\cdot\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma-1}\\
&\leq 2\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1\cdot\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma-1}\\
&\leq 2(d+1)^{1/2}\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma-1}\cdot\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2\\
&\leq 2(d+1)^{1/2}\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{2(\gamma-1)}\right)^{1/2}\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2^2\right)^{1/2}.
\end{aligned}
$$

For the second term, note that $\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1=0$ for $g\in\mathcal{A}_{i2}$. Thus, the second term is less than or equal to zero. Combining above results and (6.5), we have

$$
\begin{aligned}
&\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2\\
&\leq\lambda_n\left|\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right|+\lambda_n\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2^{\gamma}\right)\\
&\leq\lambda_n\left|\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{\gamma}-\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right|\\
&\leq 2\lambda_n\phi_n\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}-\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_2^2\right)^{1/2}
\end{aligned}
$$

It follows that

$$
\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2\leq 2\lambda_n\phi_n(d+1)^{1/2}\|\widetilde{\boldsymbol{\alpha}}^0-\widehat{\boldsymbol{\alpha}}\|_2, \tag{6.6}
$$

where $\phi_n=(d+1)^{1/2}\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}}\|\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}}\|_1^{2(\gamma-1)}\right)^{1/2}$.

On the other hand, straightforward calculation gives that

$$
\begin{aligned}
\|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2-\|\boldsymbol{y}-\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 &= (\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}\widehat{\boldsymbol{\alpha}}-(\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0)^T\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0-2\boldsymbol{y}^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\\
&= (\boldsymbol{U}\widehat{\boldsymbol{\alpha}}+\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0-2\boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0-2\boldsymbol{\epsilon}_*)^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\\
&= \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\|_2^2-2\boldsymbol{\epsilon}_*^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\\
&\geq \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\|_2^2-2\left|\boldsymbol{\epsilon}_*^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\right|\\
&\geq \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widetilde{\boldsymbol{\alpha}}^0)\|_2^2-2\sum_{i=1}^{p}\left|\boldsymbol{\epsilon}_*^T\boldsymbol{U}^{(i)}(\widehat{\boldsymbol{\alpha}}_i-\widetilde{\boldsymbol{\alpha}}_i^0)\right|,
\end{aligned}
$$

where $\boldsymbol{U}^{(i)}$ is the columns of $\boldsymbol{U}$ corresponding to $\boldsymbol{\alpha}_i$, and $\boldsymbol{\epsilon}_* = \boldsymbol{\epsilon} - \boldsymbol{e}$, $\boldsymbol{e} = (\boldsymbol{e}_1^T, \ldots, \boldsymbol{e}_n^T)^T$ with

$$\boldsymbol{e}_k = (\mathcal{U}_k(t_{k1})\widetilde{\boldsymbol{\alpha}}^0 - \boldsymbol{x}_k(t_{k1})^T\boldsymbol{\beta}(t_{k1}), \ldots, \mathcal{U}_k(t_{kn_k})\widetilde{\boldsymbol{\alpha}}^0 - \boldsymbol{x}_k(t_{kn_k})^T\boldsymbol{\beta}(t_{kn_k}))^T .$$

Let $\delta_{ni} = \|\widehat{\boldsymbol{\alpha}}_i - \widetilde{\boldsymbol{\alpha}}_i^0\|_2$, then by Lemma 4′, $\|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0)\|_2^2 \geq C_1 n K_n^{-1} \sum_i \delta_{ni}^2$. In addition, applying Cauchy-Schwarz inequality yields that $|\boldsymbol{\epsilon}_*^T\boldsymbol{U}^{(i)}(\widehat{\boldsymbol{\alpha}}_i - \widetilde{\boldsymbol{\alpha}}_i^0)|^2 \leq \delta_{ni}^2(\boldsymbol{\epsilon}_*^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{\epsilon}_*)$. Further,

$$E(\boldsymbol{\epsilon}_*^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{\epsilon}_*) = E(\boldsymbol{\epsilon}^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{\epsilon}) + E(\boldsymbol{e}^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{e}). \tag{6.7}$$

The first term on the right hand side of (6.7) equals

$$E(\boldsymbol{\epsilon}^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{\epsilon}) = E\left[\left(\sum_{k=1}^n \boldsymbol{\epsilon}_k^T\boldsymbol{U}_k^{(i)}\right)\left(\sum_{k=1}^n \boldsymbol{U}_k^{(i)T}\boldsymbol{\epsilon}_k\right)\right] = \sum_{k=1}^n E\left(\boldsymbol{\epsilon}_k^T\boldsymbol{U}_k^{(i)}\boldsymbol{U}_k^{(i)T}\boldsymbol{\epsilon}_k\right)$$

where $\boldsymbol{\epsilon}_k = (\epsilon(t_{k1}), \ldots, \epsilon(t_{kn_k}))^T$ and $\boldsymbol{U}_k^{(i)} = (x_i(t_{k1})\boldsymbol{B}_i(t_{k1}), \ldots, x_i(t_{kn_k})\boldsymbol{B}_i(t_{kn_k}))^T$.

By B-spline property and (A1), we have

$$E(B_{ij}(t_{kl})B_{ij}(t_{kl'})) = E(B_{ij}(t_{kl}))E(B_{ij}(t_{kl'})) \leq D_1^2/K_i^2,$$

and

$$E(B_{ij}^2(t)) \leq E(B_{ij}(t)) \leq D_1/K_i$$

and for some positive constant D_1. Then, by (A3), (A4) and (A5), we have

$$\begin{aligned}
& E\left(\boldsymbol{\epsilon}_k^T\boldsymbol{U}_k^{(i)}\boldsymbol{U}_k^{(i)T}\boldsymbol{\epsilon}_k\right) \\
= \;& E\left[\left(\sum_{l=1}^{n_k} \epsilon_k(t_{kl})x_i(t_{kl})\boldsymbol{B}_i^T(t_{kl})\right)\left(\sum_{l=1}^{n_k} \epsilon_k(t_{kl})x_i(t_{kl})\boldsymbol{B}_i(t_{kl})\right)\right] \\
= \;& E\left[\sum_l\sum_j \epsilon_k^2(t_{kl})x_i^2(t_{kl})B_{ij}^2(t_{kl}) + \sum_{l\neq l'}\sum_j \epsilon_k(t_{kl})x_i(t_{kl})B_{ij}(t_{kl})\epsilon_k(t_{kl'})x_i(t_{kl'})B_{ij}(t_{kl'})\right] \\
\leq \;& D_2(n_k + n_k(n_k-1)K_n^{-1})M_4M_3^2
\end{aligned}$$

for some positive constant D_2. Therefore,

$$E(\boldsymbol{\epsilon}^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{\epsilon}) \leq \sum_{k=1}^n (n_k + n_k(n_k-1)K_n^{-1})M_4M_3^2D_2 = O(n)$$

with uniformly bounded n_k, $k = 1, \ldots, n$.

The second term on the right hand side of (6.7) equals

$$E(\boldsymbol{e}^T \boldsymbol{U}^{(i)} \boldsymbol{U}^{(i)T} \boldsymbol{e}) = \sum_{k=1}^{n} E(\boldsymbol{e}_k^T \boldsymbol{U}_k^{(i)} \boldsymbol{U}_k^{(i)T} \boldsymbol{e}_k) + \sum_{k \neq k'} E(\boldsymbol{e}_k^T \boldsymbol{U}_k^{(i)} \boldsymbol{U}_{k'}^{(i)T} \boldsymbol{e}_{k'})$$

where $\boldsymbol{U}_k^{(i)} = (x_i(t_{k1})\boldsymbol{B}_i(t_{k1}), \ldots, x_i(t_{kn_k})\boldsymbol{B}_i(t_{kn_k}))^T$.

Suppose that the first p_0 covariates are the relevent ones, then $\beta_i = \widetilde{\beta}_i^0 = 0$ for $i = p_0+1, \ldots, p$. And thus, $\sum_{i=1}^{p} \|\beta_i - \widetilde{\beta}_i^0\|_\infty = \sum_{i=1}^{p_0} \|\beta_i - \widetilde{\beta}_i^0\|_\infty$. Morover, $\|\beta_i - \widetilde{\beta}_i^0\|_\infty \leq C_0 K_i^{-2}$ for some positive constant C_0 by Lemma 1 of Wang and Kai [2015]. Then, by (A3) and (A5), we have

$$\begin{aligned} |e_k(t_{kl})| &= |\mathcal{U}_k(t_{kl})\widetilde{\boldsymbol{\alpha}}^0 - \boldsymbol{x}_k(t_{kl})^T \boldsymbol{\beta}(t_{kl})| = \left| \sum_{i=1}^{p} x_i(t_{kl})(\widetilde{\beta}_i^0(t_{kl}) - \beta_i(t_{kl})) \right| \\ &\leq M_3 \sum_{i=1}^{p} \|\beta_i - \widetilde{\beta}_i^0\|_\infty = M_3 \sum_{i=1}^{p_0} \|\beta_i - \widetilde{\beta}_i^0\|_\infty \leq p_0 C K_n^{-2}, \end{aligned}$$

for some positive constant C. Thus, we have $E(e(t_{kl})e(t_{k'l'})) \leq O(K_n^{-4})$ for all k, k', l and l'.

Consequently, we have

$$\begin{aligned} & E(\boldsymbol{e}_k^T \boldsymbol{U}_k^{(i)} \boldsymbol{U}_k^{(i)T} \boldsymbol{e}_k) \\ = \; & E\left[\left(\sum_{l=1}^{n_k} e_k(t_{kl}) x_i(t_{kl}) \boldsymbol{B}_i^T(t_{kl}) \right) \left(\sum_{l=1}^{n_k} e_k(t_{kl}) x_i(t_{kl}) \boldsymbol{B}_i(t_{kl}) \right) \right] \\ = \; & E\left[\sum_{l} \sum_{j} e_k^2(t_{kl}) x_i^2(t_{kl}) B_{ij}^2(t_{kl}) + \sum_{l \neq l'} \sum_{j} e_k(t_{kl}) x_i(t_{kl}) B_{ij}(t_{kl}) e_k(t_{kl'}) x_i(t_{kl'}) B_{ij}(t_{kl'}) \right] \\ \leq \; & O([n_k + n_k(n_k-1)K_n^{-1}]K_n^{-4}) = O(K_n^{-4}), \end{aligned}$$

and

$$\begin{aligned} & E(\boldsymbol{e}_k^T \boldsymbol{U}_k^{(i)} \boldsymbol{U}_{k'}^{(i)T} \boldsymbol{e}_{k'}) \\ = \; & E\left[\left(\sum_{l=1}^{n_k} e_k(t_{kl}) x_i(t_{kl}) \boldsymbol{B}_i^T(t_{kl}) \right) \left(\sum_{l'=1}^{n_{k'}} e_{k'}(t_{k'l'}) x_i(t_{k'l'}) \boldsymbol{B}_i(t_{k'l'}) \right) \right] \\ \leq \; & O(n_k n_{k'} K_n^{-1} K_n^{-4}) = O(K_n^{-5}). \end{aligned}$$

Therefore, $E(\boldsymbol{e}^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{e}) \leq O(\sum_{k=1}^n K_n^{-4} + \sum_{k\neq k'} K_n^{-5}) = O(n)$, and then (6.7) yields $E(\boldsymbol{\epsilon}_*^T\boldsymbol{U}^{(i)}\boldsymbol{U}^{(i)T}\boldsymbol{\epsilon}_*) = O(n)$. Thus, we have

$$\|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2 - \|\boldsymbol{y} - \boldsymbol{U}\widetilde{\boldsymbol{\alpha}}^0\|_2^2 \geq O_p(nK_n^{-1})\sum_{i=1}^p \delta_{ni}^2 - O_p(n^{1/2})\sum_{i=1}^p \delta_{ni}. \tag{6.8}$$

Combining (6.6) and (6.8), we have

$$O_p(nK_n^{-1})\sum_{i=1}^p \delta_{ni}^2 - O_p(n^{1/2})\sum_{i=1}^p \delta_{ni} \leq 2\lambda_n\phi_n(d+1)^{1/2}(\sum_{i=1}^p \delta_{ni}^2)^{1/2}.$$

We further note that $\sum_{i=1}^p \delta_{ni} \leq p^{1/2}(\sum_{i=1}^p \delta_{ni}^2)^{1/2}$, and then we have

$$O_p(nK_n^{-1})\sum_{i=1}^p \delta_{ni}^2 \leq O_p(n^{1/2}p^{1/2} + \lambda_n\phi_n)(\sum_{i=1}^p \delta_{ni}^2)^{1/2}$$

By assumption (S1'), we have $\|\widehat{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\alpha}}^0\|_2^2 = O_p(n^{-1}K_n^2 p)$.

□

Proof of sparsistency in Theorem 3. First, for any i, define $\widehat{\alpha}_{ij}^*$ in the following way. Let $\widehat{\alpha}_{ij}^* = 0$ if $\{j-d,\ldots,j\} \cap \mathcal{A}_{i2} \neq \emptyset$, otherwise, $\widehat{\alpha}_{ij}^* = \widehat{\alpha}_{ij}$. Note that $\widehat{\boldsymbol{\alpha}}_{A_{ig}}^* = \boldsymbol{0}$ for $g \in \mathcal{A}_{i2}$.

By Karush-Kuhn-Tucker conditions, for $\widehat{\alpha}_{ij} \neq 0$ we have

$$2(\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T U_{(ij)} = \sum_{g=j-d}^{j} \gamma\lambda_n \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}\mathrm{sgn}(\widehat{\alpha}_{ij}),$$

where $U_{(ij)}$ is the column of $\boldsymbol{U}$ corresponding to $\widehat{\alpha}_{ij}$. Multiplying both sides by $(\widehat{\alpha}_{ij} - \widehat{\alpha}_{ij}^*)$ yields

$$\begin{aligned}
2(\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}^*) &= \sum_{i,j}\sum_{g=j-d}^{j} \gamma\lambda_n \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}\mathrm{sgn}(\widehat{\alpha}_{ij})(\widehat{\alpha}_{ij} - \widehat{\alpha}_{ij}^*) \\
&= \gamma\lambda_n \sum_{i,j} \sum_{g\in\mathcal{A}_{i2}\cap\{j-d,\ldots,j\}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}|\widehat{\alpha}_{ij}| \cdot \\
&= \gamma\lambda_n \sum_{i=1}^p \sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}(\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1 - \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}^*\|_1).
\end{aligned}$$

Note that, $(\widehat{\alpha}_{ij} - \widehat{\alpha}_{ij}^*)\mathrm{sgn}(\widehat{\alpha}_{ij}) = |\widehat{\alpha}_{ij}|$ if $\{j-d,\ldots,j\} \cap \mathcal{A}_{i2} \neq \emptyset$.

Since $\gamma b^{\gamma-1}(b-a) \leq b^\gamma - a^\gamma$ for $0 \leq a \leq b$, we have, for $g \in \mathcal{A}_{i1}$,

$$\gamma\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma-1}(\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1 - \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}^*\|_1) \leq \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma - \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}^*\|_1^\gamma.$$

Consequently, we have

$$2\left|(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*)\right| \le \lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}} (\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma - \|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^\gamma) + \gamma\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma. \quad (6.9)$$

By the minimality of $\widehat{\boldsymbol{\alpha}}$, we have

$$\lambda_n \sum_{i=1}^{p}\sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma - \lambda_n \sum_{i=1}^{p}\sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^\gamma \le \|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}^*\|_2^2 - \|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2.$$

Since $\|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1 = 0$ for $g \in \mathcal{A}_{i2}$, we have

$$\sum_{i=1}^{p}\sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^\gamma = \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}} \|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^\gamma,$$

and

$$\begin{aligned}
&2\left|(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*)\right| + (1-\gamma)\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma \\
&\le \lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}} (\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma - \|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^\gamma) + \lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma \\
&= \lambda_n \sum_{i=1}^{p}\sum_{g=1}^{G_i} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma - \lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i1}} \|\widehat{\boldsymbol{\alpha}}^*_{A_{ig}}\|_1^\gamma \\
&\le \|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}^*\|_2^2 - \|\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2 \\
&= \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}^*-\widehat{\boldsymbol{\alpha}})\|_2^2 + 2(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*).
\end{aligned}$$

By Lemma 4′ we have

$$\begin{aligned}
&(1-\gamma)\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma \\
&\le \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}^*-\widehat{\boldsymbol{\alpha}})\|_2^2 + 2(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*) - 2\left|(\boldsymbol{y}-\boldsymbol{U}\widehat{\boldsymbol{\alpha}})^T\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}-\widehat{\boldsymbol{\alpha}}^*)\right| \\
&\le \|\boldsymbol{U}(\widehat{\boldsymbol{\alpha}}^*-\widehat{\boldsymbol{\alpha}})\|_2^2 \\
&\le nC_2K_n^{-1}\sum_{i=1}^{p}\|\widehat{\boldsymbol{\alpha}}_i^*-\widehat{\boldsymbol{\alpha}}_i\|_2^2
\end{aligned}$$

Since $\widetilde{\boldsymbol{\alpha}}^0_{A_{ig}} = \boldsymbol{0}$ for $g \in \mathcal{A}_{i2}$, we have $\|\widehat{\boldsymbol{\alpha}}_i^*-\widehat{\boldsymbol{\alpha}}_i\|_2^2 \le \|\widehat{\boldsymbol{\alpha}}_i-\widetilde{\boldsymbol{\alpha}}_i^0\|_2^2$, and

$$(1-\gamma)\lambda_n \sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}} \|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^\gamma \le nC_2K_n^{-1}\sum_{i=1}^{p}\|\widehat{\boldsymbol{\alpha}}_i-\widetilde{\boldsymbol{\alpha}}_i^0\|_2^2 = O_p(pK_n)$$

Since

$$\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\geq\left(\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1\right)^{\gamma}\geq\|\widehat{\boldsymbol{\alpha}}^{*}-\widehat{\boldsymbol{\alpha}}\|_2^{\gamma},$$

then if $\|\widehat{\boldsymbol{\alpha}}^{*}-\widehat{\boldsymbol{\alpha}}\|_2^2=\sum_{i=1}^{p}\|\widehat{\boldsymbol{\alpha}}_i^{*}-\widehat{\boldsymbol{\alpha}}_i\|_2^2>0$, we have

$$\begin{aligned}(1-\gamma)\lambda_n &\leq nC_2K_n^{-1}\sum_{i=1}^{p}\|\widehat{\boldsymbol{\alpha}}_i^{*}-\widehat{\boldsymbol{\alpha}}_i\|_2^2\left\{\sum_{i=1}^{p}\sum_{g\in\mathcal{A}_{i2}}\|\widehat{\boldsymbol{\alpha}}_{A_{ig}}\|_1^{\gamma}\right\}^{-1}\\ &\leq nC_2K_n^{-1}\|\widehat{\boldsymbol{\alpha}}^{*}-\widehat{\boldsymbol{\alpha}}\|_2^{2-\gamma}\\ &\leq O_p(n^{\gamma/2}K_n^{1-\gamma}p^{1-\gamma/2}),\end{aligned}$$

and thus

$$\Pr\left\{\|\widehat{\boldsymbol{\alpha}}^{*}-\widehat{\boldsymbol{\alpha}}\|_2^2>0\right\}\leq\Pr\left\{\frac{\lambda_n}{n^{\gamma/2}K_n^{1-\gamma}p^{1-\gamma/2}}\leq O_p(1)\right\}.$$

By assumption (S2'), the right hand side converges to zero as n goes to infinity, which implies that $(\widehat{\boldsymbol{\alpha}}_{A_{ig}}:g\in\mathcal{A}_{i2})=\mathbf{0}$ with probability approaching to 1. □

CHAPTER 7

Supplementary Material

In this part, we provide additional simulation results for the case where the coefficient functions live in the linear space spanned by B-spline basis functions.

For each scenario in Chapter 4, we conducted additional simulations in which the nonzero coefficient functions belong to B-spline spaces. Our new simulation studies are called Scenarios 1B, 2B.1, 2B.2 and 3B corresponding to the settings of Scenarios 1, 2.1, 2.2 and 3. Note that, in each new scenario, there are four nonzero coefficient functions, shown in Figure 7.1. In addition, we have

- Scenario 1B: $n = 200$ and $p = 5$
- Scenario 2B.1: $n = 200$ and $p = 20$
- Scenario 2B.2: $n = 400$ and $p = 20$
- Scenario 3B: $n = 200$ and $p = 1000$.

MISE values are summarized in Tables 7.1, 7.3, 7.5 and 7.7 and the measures of functional sparsity are summarized in Tables 7.2, 7.4,7.6 and 7.8. The last column of MISE tables for Scenarios 2B and 3B gives the maximum MISE values among the coefficient functions with global sparsity. Bias plots of nonzero coefficient functions are displayed in Figures 7.2, 7.5, 7.7 and 7.9. The asymptotic standard errors are displayed in Figures 7.3, 7.6, 7.8 and 7.10. It can be seen that all results are in accordance with the general conclusion drawn in Chapter 4.

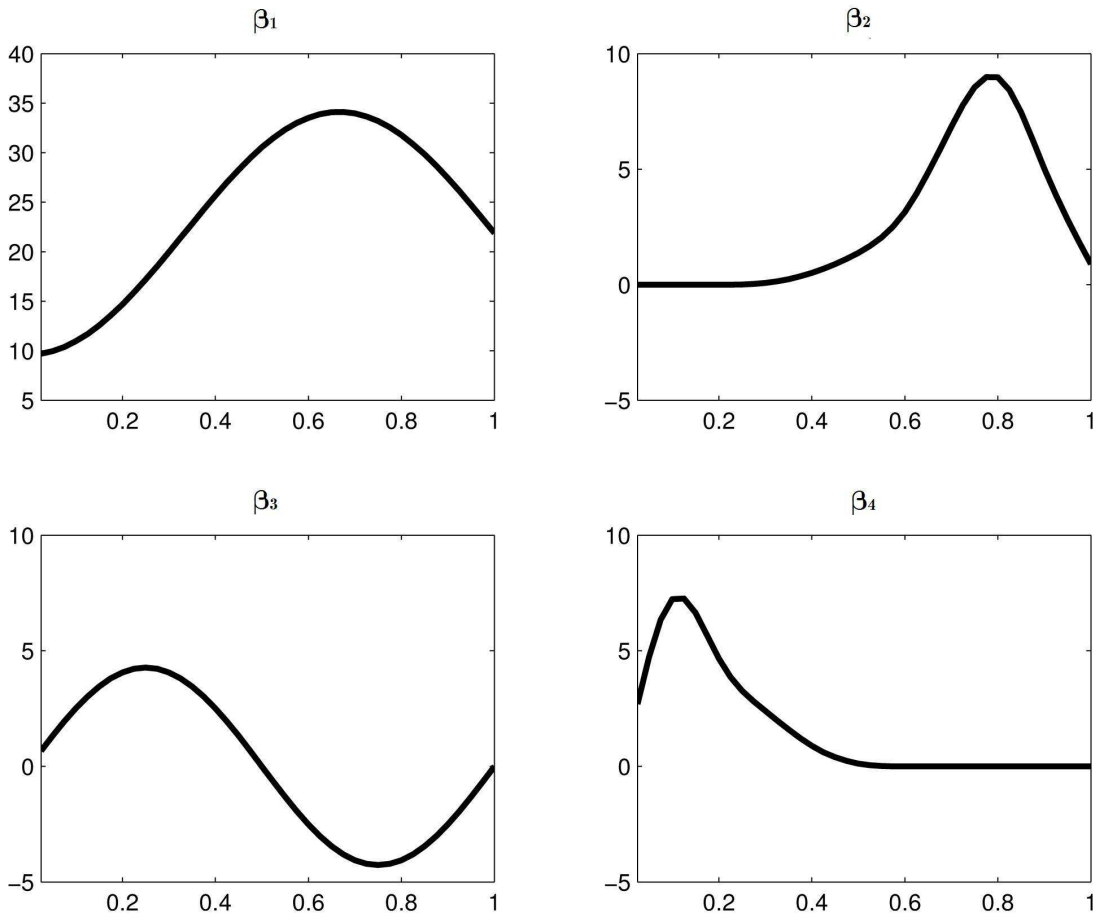

Figure 7.1: A graphical illustration of the nonzero coefficient functions β_i $i = 1, \ldots, 4$ in Scenarios 1B, 2B.1, 2B.2, and 3B.

Table 7.1: Comparison of MISE for each coefficient function in Scenario 1B.

Method	MISE				
	β_1	β_2	β_3	β_4	β_5
LSE	0.3680	0.0297	0.0096	0.0319	0.3184
Lasso	1.0717	0.0584	0.0092	0.0209	0.0041
$\text{PLSE}_{0.5}$	0.3906	0.0234	0.0086	0.0170	0

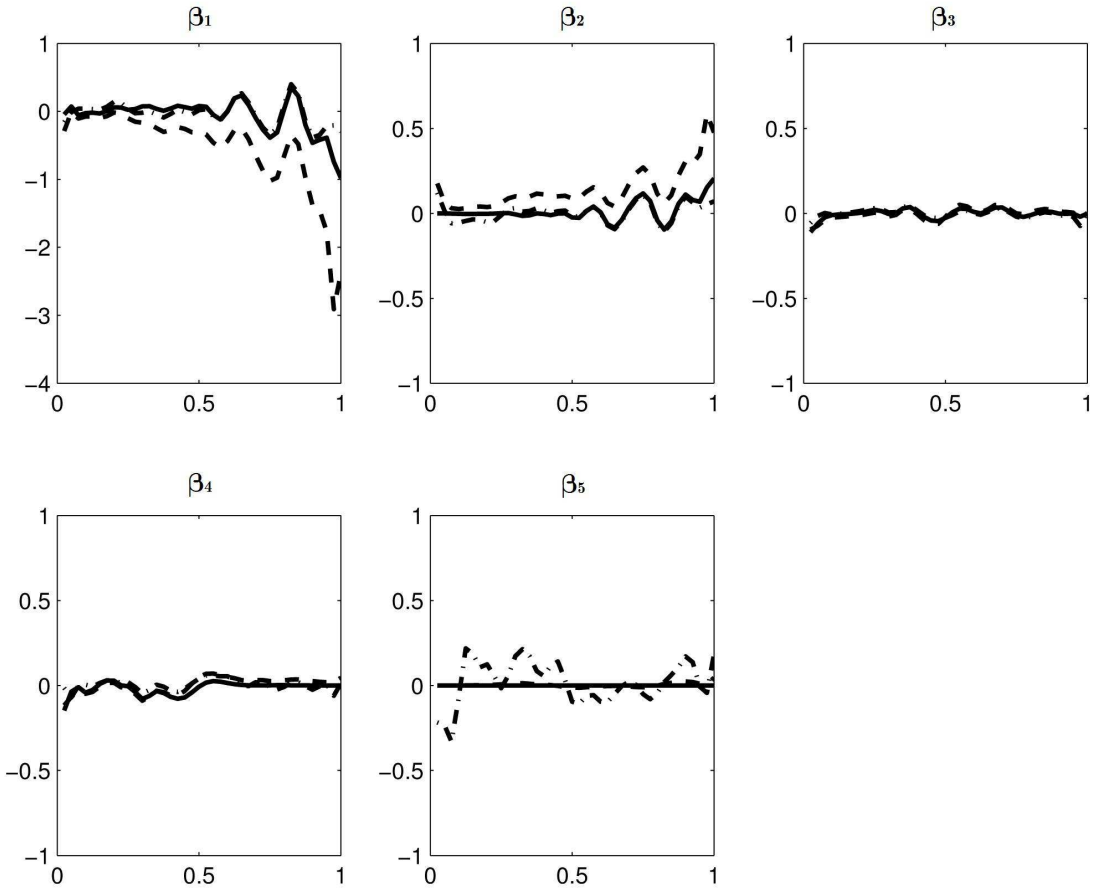

Figure 7.2: Estimation bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 1B.

Table 7.2: Sparsity summary measures (a)-(d) in Scenario 1B. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 5$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	$C_{5,0}$	$I_{5,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.018	0.002	0	0.010	0.062	0.003	0.852	0	0.477	0
$\text{PLSE}_{0.5}$	0	0	0.159	0.025	0	0.019	0.370	0.027	1	0	1	0
true model	0	1	0.2	0.8	0	1	0.425	0.575	1	0	1	4

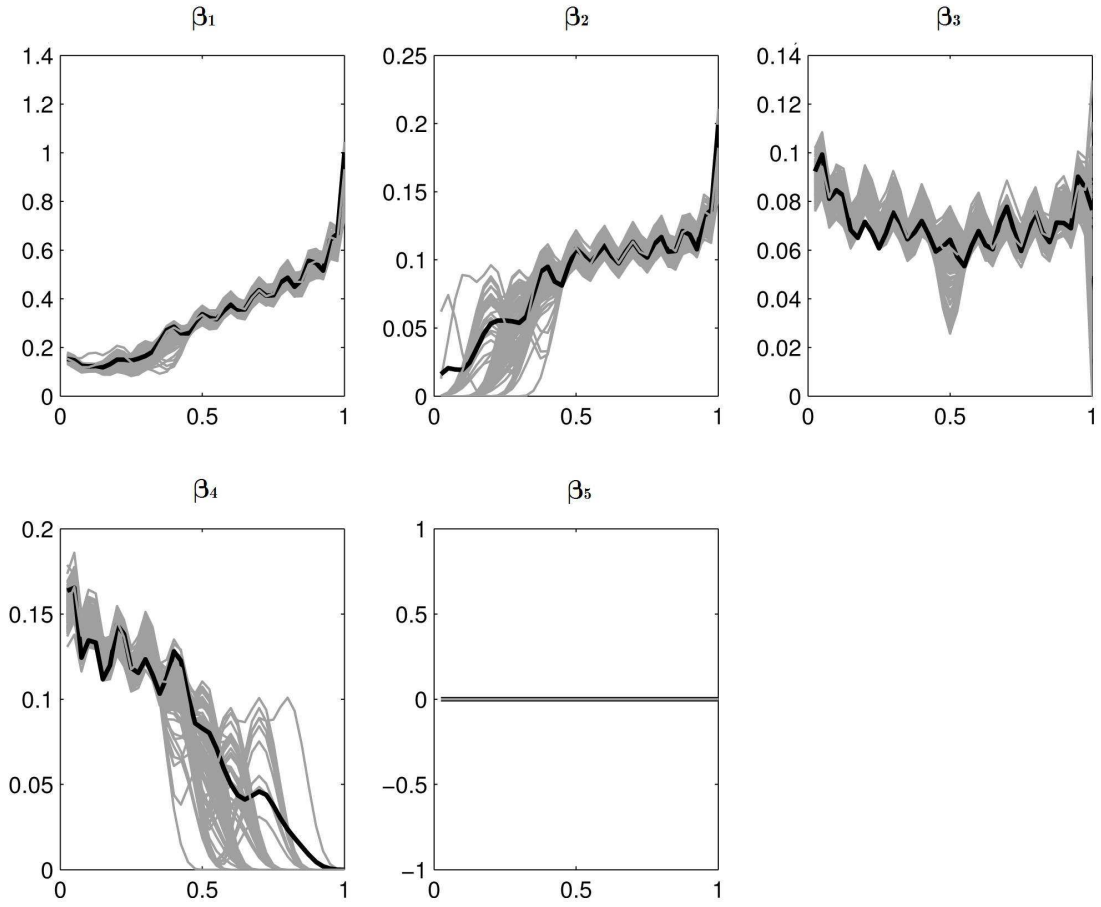

Figure 7.3: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions with fixed number of knots in Scenario 1B.

Table 7.3: Comparison of MISE for each coefficient function in Scenario 2B.1.

Method	MISE				$\max_{i\geq 5}\text{MISE}_i$
	β_1	β_2	β_3	β_4	
LSE	0.1949	0.0163	0.0074	0.0227	0.2474
Lasso	8.3178	0.3879	0.0127	0.0167	0.0006
$\text{PLSE}_{0.5}$	0.2926	0.0146	0.0059	0.0106	0.0000

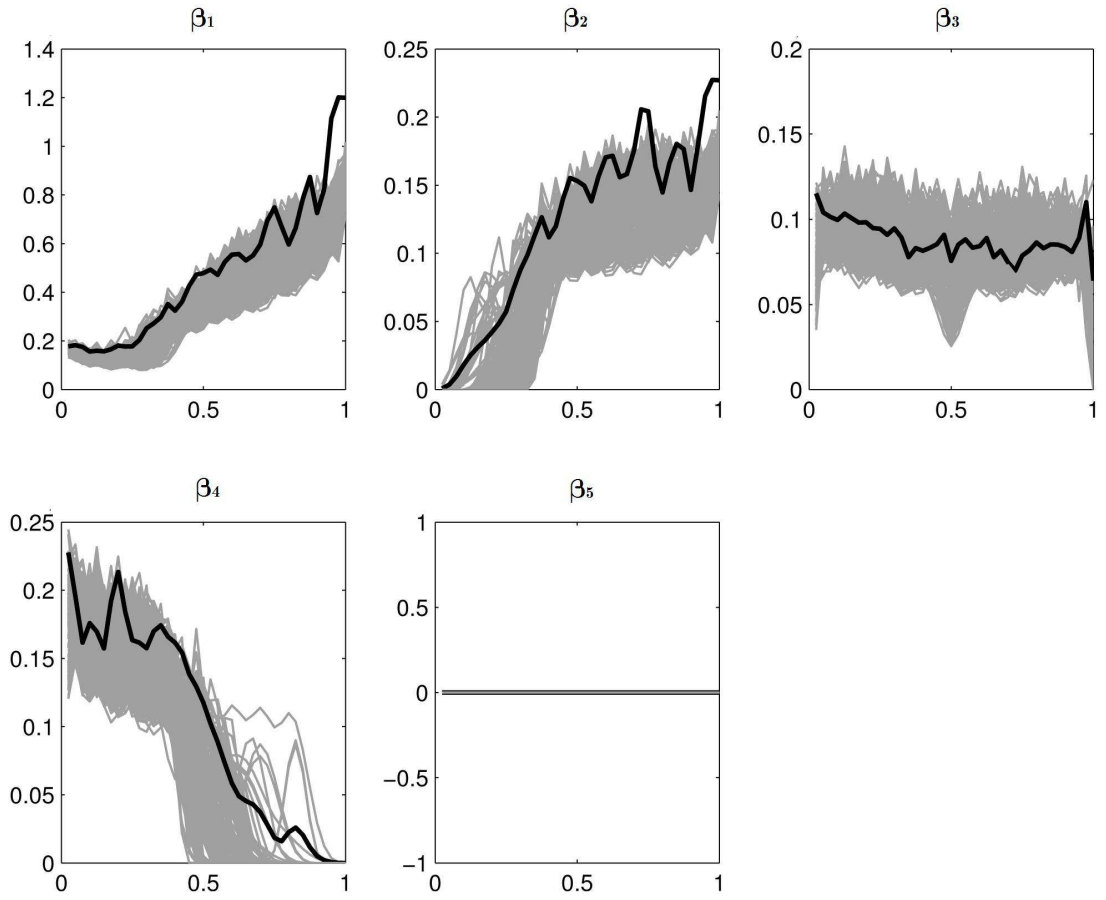

Figure 7.4: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions with adaptive number of knots in Scenario 1B.

Table 7.4: Sparsity summary measures (a)-(d) in Scenario 2B.1. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.009	0	0	0.022	0.282	0	3.385	0
$\text{PLSE}_{0.5}$	0	0	0.189	0.001	0	0.024	0.392	0	15.89	0
true model	0	1	0.2	0.8	0	1	0.425	0.575	16	4

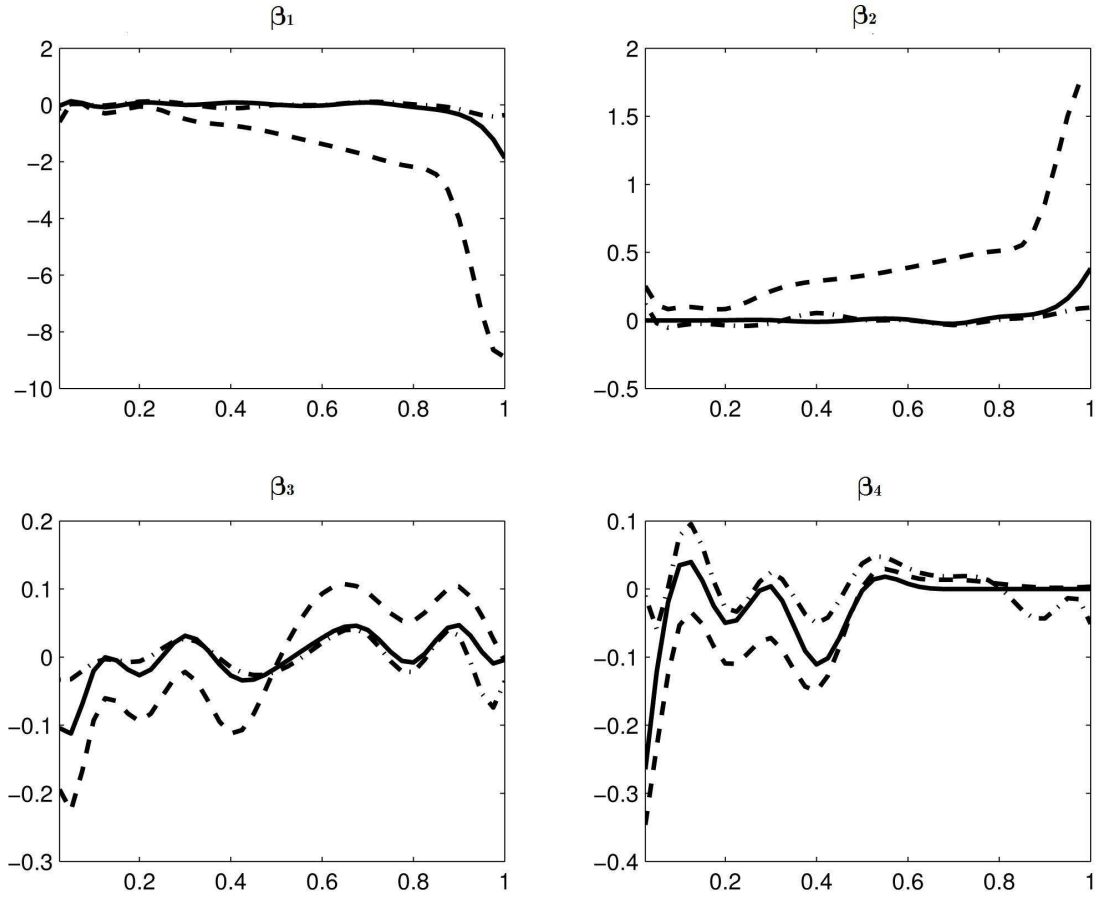

Figure 7.5: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 2B.1.

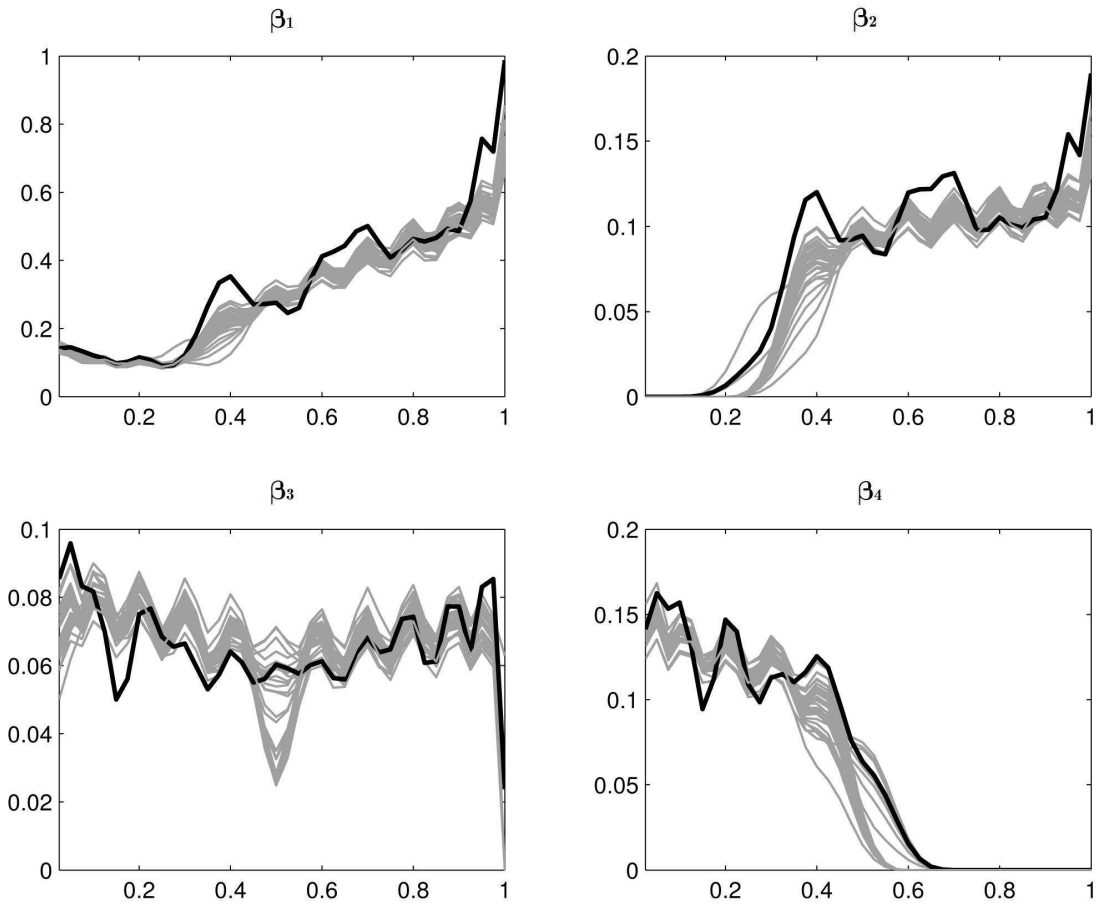

Figure 7.6: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions in Scenario 2B.1.

Table 7.5: Comparison of MISE for each coefficient function in Scenario 2B.2.

Method	MISE				$\max_{i\geq 5}$ MISE$_i$
	β_1	β_2	β_3	β_4	
LSE	0.0762	0.0064	0.0027	0.0087	0.0914
Lasso	2.2156	0.1022	0.0047	0.0058	0.0004
PLSE$_{0.5}$	0.0752	0.0044	0.0022	0.0038	0.0000

Table 7.6: Sparsity summary measures (a)-(d) in Scenario 2B.2. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.01	0	0	0.019	0.29	0	2.11	0
PLSE$_{0.5}$	0	0	0.164	0	0	0.018	0.39	0	15.81	0
true model	0	1	0.2	0.8	0	1	0.425	0.575	16	4

Table 7.7: Comparison of MISE for each coefficient function in Scenario 3B.

Method	MISE				$\max_{i\geq 5}$ MISE$_i$
	β_1	β_2	β_3	β_4	
LSE	0.1916	0.0164	0.0065	0.0211	0.0058
Lasso	3.2495	0.1536	0.0077	0.0128	0.0012
PLSE$_{0.5}$	0.2584	0.0134	0.0055	0.0101	0.0001

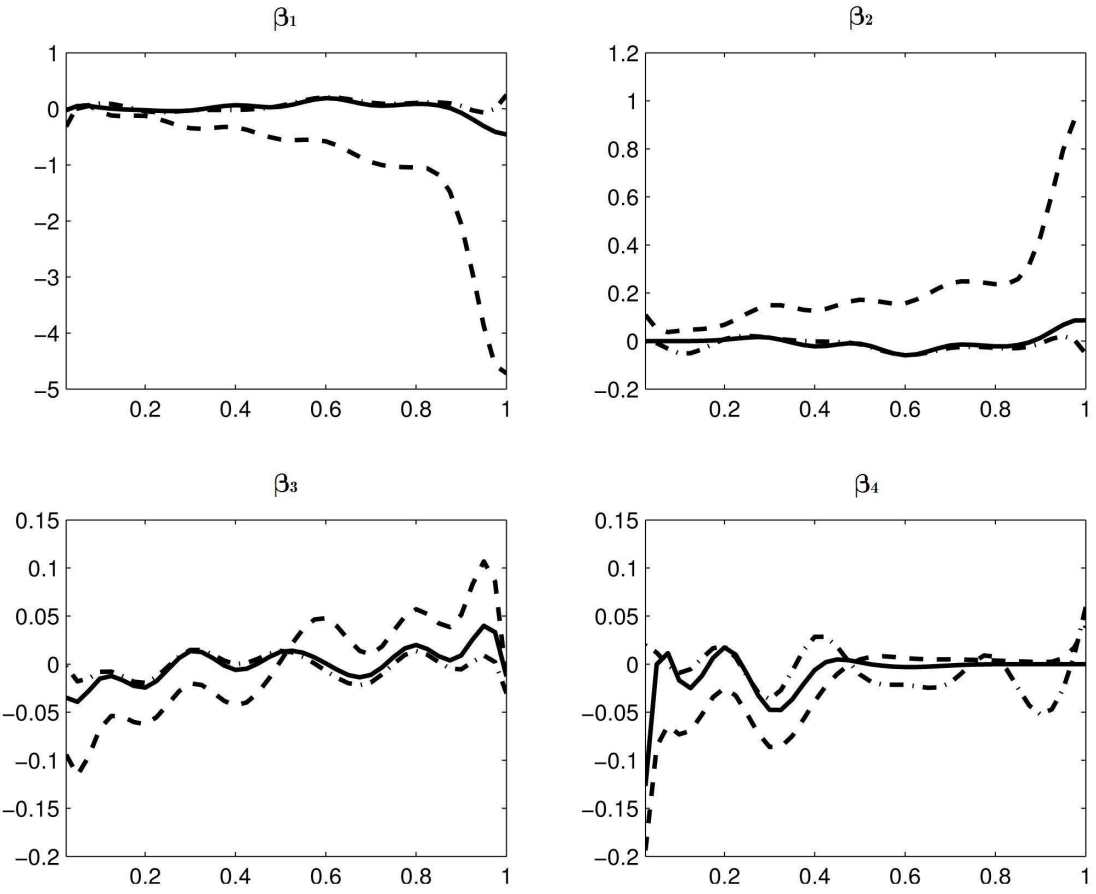

Figure 7.7: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 2B.2.

Table 7.8: Sparsity summary measures (a)-(d) in Scenario 3B. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	C_0	I_0
LSE	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.016	0	0	0.018	0.222	0	0.1	0
$\text{PLSE}_{0.5}$	0	0	0.181	0.001	0	0.022	0.391	0.001	4.86	0
true model	0	1	0.2	0.8	0	1	0.425	0.575	5	4

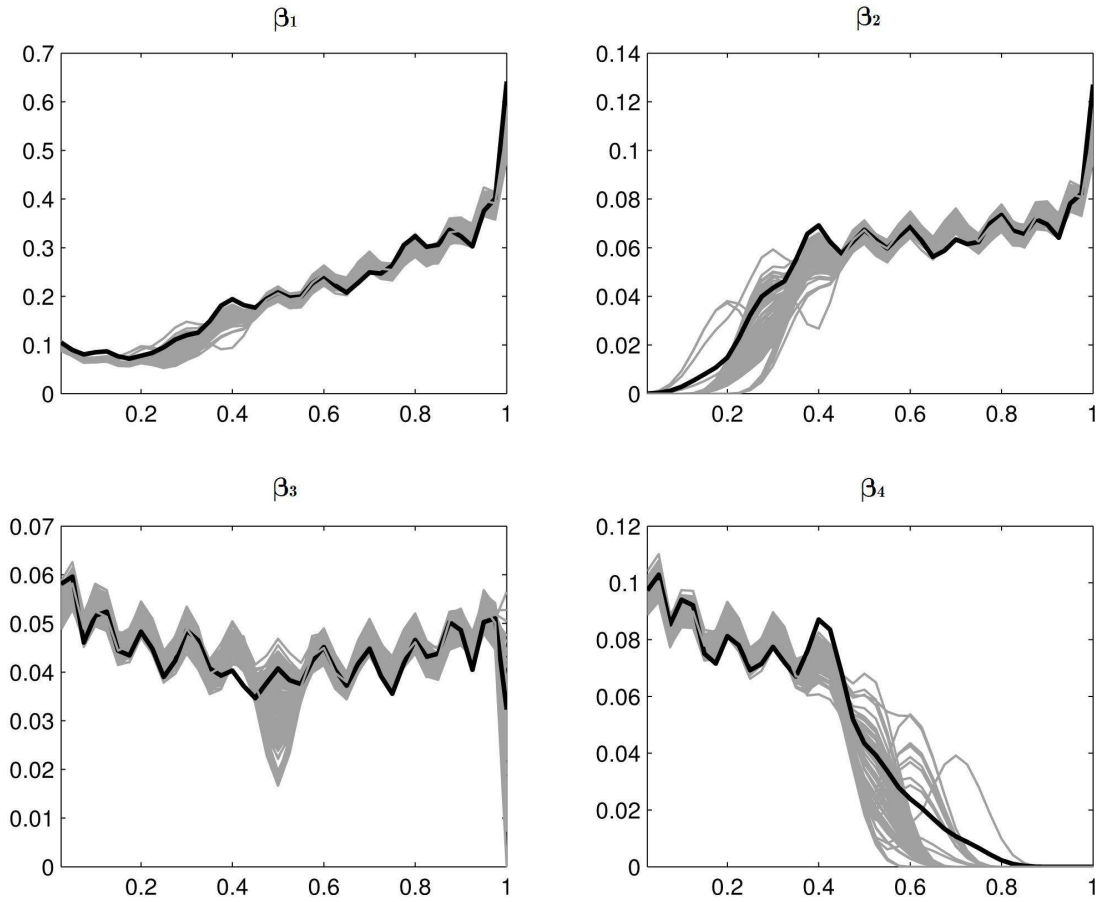

Figure 7.8: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions in Scenario 2B.2.

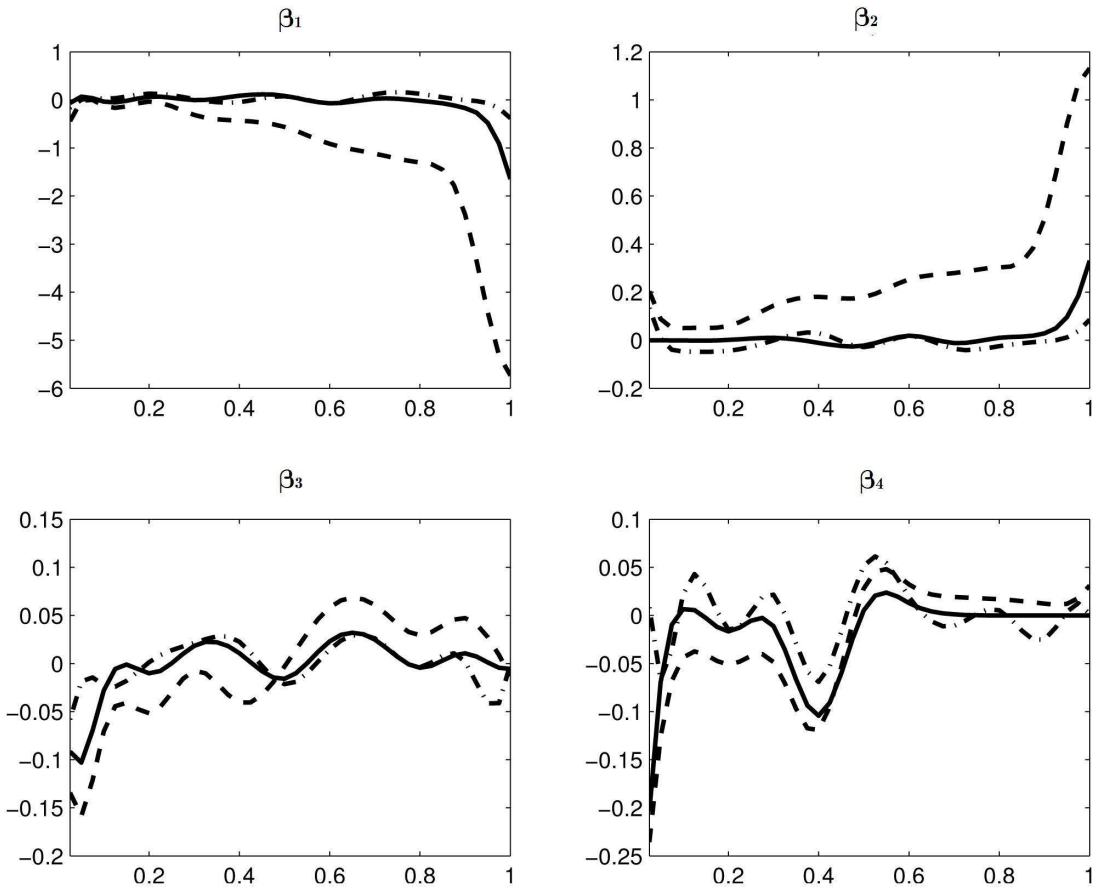

Figure 7.9: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 3B.

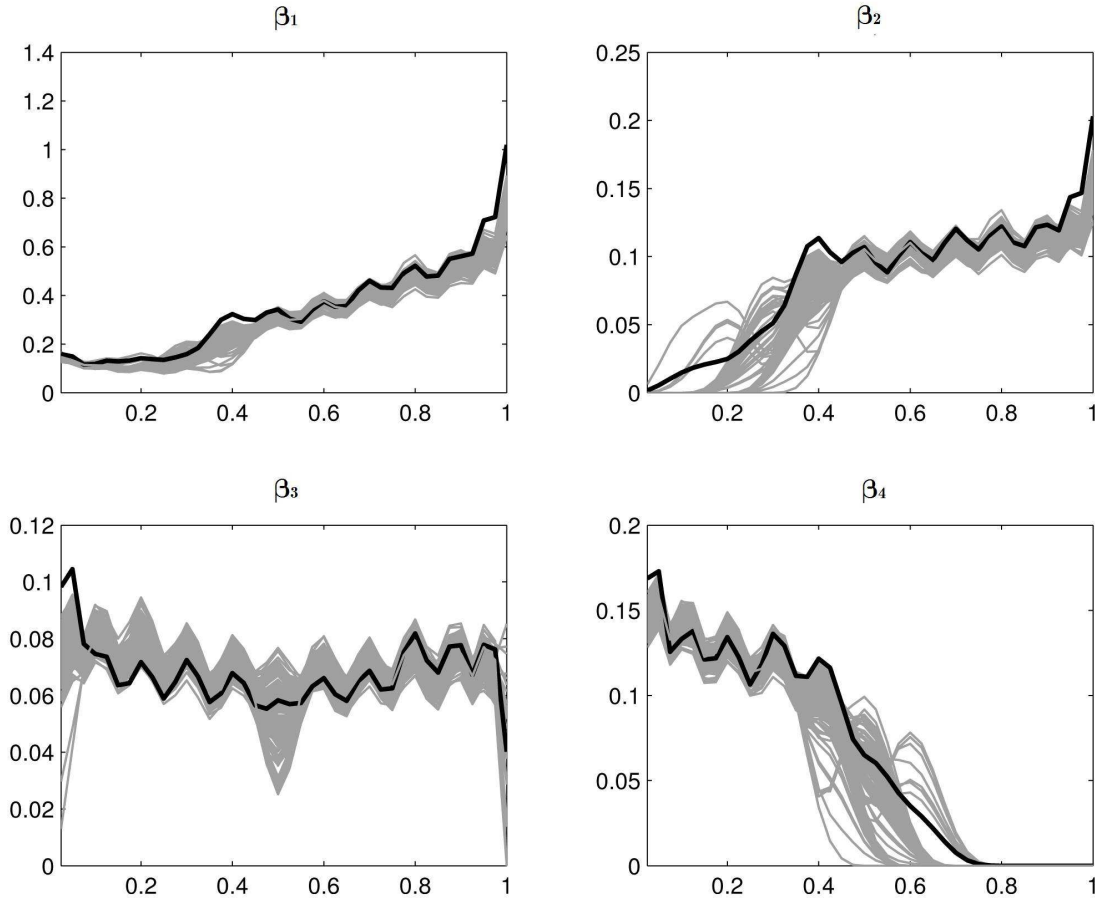

Figure 7.10: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions in Scenario 3B.

CHAPTER 8

Conclusions and Future Work

8.1 Conclusions

In this dissertation, we proposed a variable selection method to efficiently achieve globe sparsity and local sparsity simultaneously. It was developed based on the group bridge penalty and the natural properties of B-splines. In the simulation study, we compared the performance of proposed penalized method (PLSE), lasso method and ordinary least squares method (LSE). It was found that the LSE provides consistent estimation, but it had neither global sparsity nor local sparsity. The lasso method achieved functional sparsity in some degree, but cannot ensure global sparsity. However, our proposed method had outstanding performance in achieving both global sparsity and local sparsity. Moreover, as the dimension of model grew, i.e., diverging dimension and ultra-high dimension, the advantage of our proposed method became manifest.

In Chapter 5, we applied the proposed method to the gene expression data and Boston housing data, and obtained some sensible results. For gene expression data, our proposed PLSE greatly reduced the size of the model and narrowed the range of possible active transcription factors. It also indicated the active period for selected TFs. For the Boston housing data, the varying coefficient model characterized the impacts of seven interesting factors (`CRIM`, `RM`, `TAX`, `NOX`, `PTRATIO`, `AGE`, `B`) on median housing values as functions of `LSTAT`. The fitting results by PLSE suggested that some factors have no effects on people with high `LSTAT` values.

Further, we established some theoretical properties for our proposed method described in Chapter 3. In particular, we demonstrated the consistency in both functional estimation and detecting functional sparsity under mild assumptions for nonparametric regression. In addition, the convergence rate of our proposed method is the optimal rate of nonparametric

regression. We also confirmed the properties of consistency and sparsistency under diverging dimensional model.

8.2 Future Work

In Chapter 4, our proposed method was applied to varying coefficient models of ultra-high dimension. The simulation results are promising. We plan to investigate the theoretical properties of our proposed method under ultra-high dimension setting.

In Section 2.4, an asymptotic variance of the PLSE was provided; however, its consistency has not been established. In the future, we will seek consistent variance estimators for the estimated coefficient functions. In addition, we plan to derive the asymptotic distribution of the estimated functions. Moreover, we will extend the proposed method to generalized linear models with varying coefficients, and establish the consistency and sparsistency for the cases of finite dimension and diverging dimension.

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In N., P. B. and F., C., editors, *Proc. of the 2nd Int. Symp. on Information Theory*, pages 267–281.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n (with discussion). *The Annals of Statistics*, 35:2313–2351.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B*, 72:3–25.

de Boor, C. (2001). *A Practical Guide to Splines*. Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

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

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B*, 70:849–911.

Fan, J., Ma, Y., and Dai, W. (2014). Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. *Journal of the American Statistical Association*, 109:1270–1284.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38(4):2282–2313.

Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, 96:339–355.

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

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

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics*, 37:2083–2108.

Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thomson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., and Young, R. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804.

Leng, C. (2009). A simple approach for varying-coefficient model selection. *Jornal of Statistical Planning and Inference*, 139.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics*, 34:2272–2297.

Ravikumar, P., Lafferty, J., and Wasserman, L. (2009). Sparse additive models. *Journal of Royal Statistical Society: Series B*, 71:1009–1030.

Schumaker, L. (1981). *Spline Functions: Basic Theory*. Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3279.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053.

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

Tu, C. Y., Park, J., and Wang, H. (2015). Estimation of functional sparsity in nonparametric varying coefficient models with functional covariates. *Submitted to Statistica Sinica*.

Tu, C. Y., Song, D., Breidt, F. J., Berger, T. W., and Wang, H. (2012). Functional model selection for sparse binary time series with multiple input. In *Economic Time Series: Modeling and Seasonality*, pages 477–497. Chapman and Hall/CRC.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wang, H. and Kai, B. (2015). Functional sparsity: Global versus local. *Statistica Sinica*.

Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23:1486–1494.

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

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *The Journal of Machine Learning Research*, 13:1973–1998.

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

Zhou, J., Wang, N.-Y., and Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23:25–50.