An Iterative Approach for Model Selection in A Class of Semiparametric Models

Xiaochen Kou

PhD

University of York

Mathematics

May 2018

ABSTRACT

The class of single-index varying coefficient (SIVC) models is an important extension of varying coefficient models and has proved to be remarkably useful in data analysis. The model selection in such class is essential but challenging due to the complicated structure of SIVC models. In this thesis, we take on this challenge and develop a novel iterative approach for model selection in SIVC models. Based on the ideas of kernel smoothing, penalised least squares with SCAD penalty and group selection, the proposed iterative approach can simultaneously select and estimate the SIVC models. Asymptotic properties of the proposed iterative approach are also established, which justify the proposed approach theoretically. Intensive simulation studies conducted in this paper illustrate the efficiency of the proposed iterative approach. Finally, we apply the SIVC model and the proposed model selection method to an environmental set from Hong Kong and a housing dataset from Boston, both of which lead to some interesting findings.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my heartily gratitude to my supervisor, Professor Wenyang Zhang, for his continuous support, inspiration and guidance throughout my Ph.D study and research. He is genuinely a brilliant statistician and also an excellent mentor. The independent, creative and critical thinking he taught me will be the treasure for my entire life.

I would like to give thanks to Professor Degui Li and Dr. Marina Knight for being my Thesis Advisory Panel members and giving me lots of help and encouragement from the beginning. I would like to acknowledge the Department of Mathematics for all the support.

I thank my PhD friend: Dr. Yuan Ke, Mr. Jiraroj Tosasukul, and Miss. Lingling Wei for all the help, interesting discussions and chats.

Besides, I would like to thank the York Advanced Research Computing Cluster (YARCC), which dramatically speeds up the computation in my numerical studies.

Finally, I would like to give special thanks to my family:

my father Rongjian Kou, my mother Jun Zhou and my devoted wife Zhongmei Ji for all their support and love.

AUTHOR'S DECLARATION

The literature review in Chapter 2 provides some key ideas related to this thesis, which includes:

- Chapter 2.1 reviews the framework of local polynomial modelling, which is mainly from the the book: *Local Polynomial Modelling and Its Applications* by Fan, J. and Gijbels, I. (1996).

- Chapter 2.2 provides a summary of the penalised least squares approach, which is mostly based on *Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties* by Fan and Li (2001) and *Sparse High Dimensional Models in Economics* by Fan *et al.*(2011).

- Chapter 2.3 briefly introduces the Generalised Information Criterion found in *Tuning Parameter Selection in High Dimensional Penalized Likelihood* by Fan and Tang (2013).

- Chapter 2.4 contains a concise review of varying coefficient models chiefly from *Adaptive Varying-coefficient*

Linear Models by Fan *et al.*(2003) and *Statistical Methods with Varying Coefficient Models* by Fan and Zhang (2008).

The rest chapters are mainly related to my submitted paper: *An iterative approach for model selection in single-index varying coefficient models*, joint with Prof. Efang Kong and Prof. Wenyang Zhang.

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, university. All sources are acknowledged as references.

List of Contents

List of Tables

List of Figures

Figure **Page**

CHAPTER 1

INTRODUCTION

Variable selection is an important topic in statistics with wide applications in diverse disciplines, such as econometrics, epidemiology and computer science.

The traditional approaches, such as stepwise selection procedures and best subsets regression, suffer from several limitations especially when the number of potential variables is big. Apart from the expensive computational cost, stepwise regression neglects the stochastic errors in the variable selection process which leads to a somewhat poor interpretation of its theoretical properties while the best subsets regression is an unstable procedure, see Breiman (1996).

The penalised likelihood/least squares approach emerged as a promising alternative and have been well studied, as it possesses many advantages over the traditional approaches.

With an appropriate penalty function, the penalised approach would automatically select significant variables and estimate coefficients simultaneously. In the family of L_p penalised least squares, the ridge regression associated with L_2 penalty are proposed by Frank and Friedman (1993) and base on L_1 penalty, the least absolute shrinkage and selection operator (LASSO) are proposed by Tibshirani (1996, 1997). Boyd and Vandenberghe (2004) developed the proximal gradient descent (PGD) algorithm to solve LASSO and other L_1 based penalised methods. Efron *et al.*(2004) proposed an efficient algorithm, termed as least angle regression (LARS), which can be used to generate the full set of LASSO solutions with a minor modification. Yuan and Lin (2006) studied and proposed efficient algorithms for the extensions of the LASSO for selecting the grouped variables. Although LASSO enjoys considerable nice properties, it is inconsistent with variable selection as the resulting penalised estimator is biased. Zou (2006) proposed the adaptive LASSO to overcome the inconsistency of the LASSO. The smoothly clipped absolute deviation (SCAD) penalty proposed by Fan and Li (2001) also enjoys the oracle properties if the regularization parameter is appropriately chosen; namely, the resulting penalised estimators perform as well as the estimators if the true underlying model were known in advance. Furthermore, Fan and Li (2001) extended the penalised least squares to likelihood-based models and

established a unified algorithm to solve both the penalised least squares and penalised likelihood via local quadratic approximations. Hunter and Li (2005) proposed an algorithm termed minorize–maximize (MM) to optimise the penalised likelihood for a broad class of penalty functions and established the convergence and other theoretical properties of MM algorithm. Based upon local linear approximation, Zou and Li (2008) developed a one-step sparse estimation procedure for optimising the penalised likelihood which can alleviate the computational burden without losing statistical efficiency.

Much literature about the application of the penalised likelihood/least squares approach on diverse high-dimensional models has emerged in the last two decades. See Fan and Lv (2008), Fan *et al.*(2009), Bickel *et al.*(2009), Wang and Xia (2009), Stefanski *et al.*(2014), Wang, Peng and Li (2015), Fan *et al.*(2015), Li, Ke and Zhang (2015), Fan and Lv (2016), Zhang *et al.*(2016), and the references therein.

The existing literature mainly focuses on linear models, varying-coefficient models, and additive models. The pre-supposed parametric linear models may ignore the dynamic feature in the data set and often be too unrealistic to work well in analysing some complex data. Instead, varying coefficient models loosen the linear restriction and let the constant coefficients evolve with certain characteristics to describe the varying relationship between the response and covariates.

Varying coefficient models are remarkably useful in exploring the dynamic patterns of the impacts of covariates in data analysis and has gained popularity in modelling and forecasting non-linear time series, analysing functional and longitudinal data during the past decade. The substantial amount of literature includes Chen and Tsay (1993), Carroll *et al.*(1998), Kauermann and Tutz (1999), Hastie and Tibshirani (1993), Cai, Fan and Yao (2000), Cai, Fan and Li (2000), Zhang and Lee (2000), Fan and Huang (2005) and Fan and Zhang (2008). The works about the hypotheses testing of the model include Fan and Zhang (2000), Fan, Zhang and Zhang (2001) and Li and Liang (2008).

Although varying-coefficient models are defined in slightly different forms from diverse statistical contexts, a typical varying-coefficient model is assumed by most previous work that

$$Y_i = \mathbf{X}_i^\top \boldsymbol{f}(Z_i) + \epsilon_i, \tag{1.1}$$

where $(\mathbf{X}_i, Y_i)$ is the i-th observation $(1 \le i \le n)$, $Y_i \in \mathbb{R}^1$ is the response variable, $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top \in \mathbb{R}^d$ is the d-dimensional vector of covariate, $Z_i \in \mathbb{R}^1$ is often called the index which is collected from the observations, the random noise $\epsilon_i \in \mathbb{R}^1$ are independent identically defined with $E(\epsilon_i \mid \mathbf{X}_i, Z_i) = 0$ and coefficient vector $\boldsymbol{f}(\cdot) = (f_1(\cdot), \ldots, f_d(\cdot))^\top \in \mathbb{R}^d$ are the vector of unknown functions of the index. And specifically, in this def-

inition of varying-coefficient model, the index Z is assumed to be a known variable which is chosen from the covariates. Since the varying coefficient models are locally linear models, it is reasonable to employ kernel polynomial smoothing to estimate, see Hoover *et al.*(1998), Wu *et al.*(1998), Xia and Li (1999) and Fan and Zhang (1999).

However, it is often not very clear which variable should be chosen as the index in practical application when it comes to the analysis of complicated data. Instead of selecting the index variable in the light of experience, it would probably be more sensible to estimate it from the data. Fan, Yao and Cai (2003) proposed the single-index varying coefficient (SIVC) model to solve the problem by generalising the index as a linear combination of covariates. Therefore, the index is set to be $Z_i = \mathbf{X_i}^\top \boldsymbol{\beta} \in \mathbb{R}^1$, $i = 1, \ldots, n$, where the index coefficient $\boldsymbol{\beta} \in \mathbb{R}^d$ is unknown and estimated by data. Then, the SIVC model assumes that

$$Y_i = \mathbf{X}_i^\top \boldsymbol{f}(\mathbf{X_i}^\top \boldsymbol{\beta}) + \epsilon_i. \tag{1.2}$$

The value of SIVC model has gone beyond the exploration of dynamic pattern. It is also a notable approach to ameliorate the "curse of dimensionality" in nonparametric modelling, see Fan and Zhang (2008). Meanwhile, the SIVC model substantially enlarges the modelling capacity, because it assumes the index to be unknown and estimated by data.

Although the SIVC model is equipped with numerous advantages, due to its sophisticated structure, it would be difficult to obtain satisfactory estimators without model selection, especially in the high dimensional situation. In most high dimensional SIVC models, only a handful covariates significantly contribute to the response variable or index variable, and hence it is necessary to consistently obtain the estimates admit sparsity. With this in mind, selecting the significant components of the model and eliminating the irrelevant components correctly is essential.

In fact, model selection in the semi-parametric models has been extensively studied in the literature. For instance, Lin and Zhang (2003), Fan and Li (2004) and Li and Liang (2008) extend the penalised estimation methods (e.g., SCAD) to select the significant sub-model in semi-parametric models. Wang *et al.*(2008) and Wang and Xia (2009) use group selection to select the significant variables in modest dimensional varying coefficient models. More recently, Song *et al.*(2012), Cheng *et al.*(2014), Fan *et al.*(2014) and Liu *et al.*(2014), Li *et al.*(2015) apply the group penalised method to select the significant covariates and estimate the functional coefficients for the high dimensional varying coefficient models. Therefore, we are motivated to establish a more specific penalised approach that can automatically select the significant components and simultaneously estimate the relevant parameters in SIVC

model.

In this thesis, based on the ideas of kernel smoothing, penalised least squares with SCAD penalty and group selection, we proposed an iterative approach to select the significant varying coefficient $\boldsymbol{f}(\cdot)$ and the relevant direction $\boldsymbol{\beta}$ in SIVC model, thereby simplify the model used. We term this selection procedure as model selection. In the meantime, our proposed selection approach is able to detect the functional coefficients with zero derivatives, which can be used for identifying the constant coefficients. To sum up, the proposed model selection has threefold aims: variable selection, index specification and the identification of the constant coefficients. Additionally, the proposed approach also applies to the computation of the penalised estimators for SIVC.

The thesis is organized as follows. We begin in Chapter 2 with a literature review on local polynomial modelling, penalised least squares, generalised information criterion and varying coefficient models. Chapter 3 describes the SIVC model and develops an iterative procedure for the estimation of the model. This also aids as a helpful stepping stone for the demonstration of the methodology in the following chapters. In Chapter 4, we propose an iterative approach for model selection and estimation for the unknowns in the SIVC model. Chapter 5 is devoted to the selection of the bandwidth (smoothing parameter) and tuning parameters (regularization

parameters). Chapter 6 provides the asymptotic properties of the proposed model selection and lists the necessary technical conditions. The performance of the proposed model selection and estimation procedures is assessed by simulation studies in Chapter 7. In Chapter 8.1, we apply the SIVC model together with the proposed iterative procedures to analyse an environmental data set from Hong Kong. This real data analysis will explore which pollutants and environmental factors significantly affect the number of daily total hospital admissions for circulatory and respiratory problems in Hong Kong and the dynamic pattern of the impacts. In Chapter 8.2, we analyse another real data example on a Boston housing data set to explore how the collected factors affect the median value of owner-occupied homes in Boston. In Chapter 9, we gives the proof of theoretical results. In particular, the Chapter 9 are mainly from my submitted paper "An Iterative Approach for Model Selection in Single-index Varying Coefficient Models" and are the joint work with Prof. Efang Kong and Prof. Wenyang Zhang.

CHAPTER 2

LITERATURE REVIEW

In this chapter, the literature we shall review is fourfold. The first part presented in Section 2.1 is about the local polynomial modelling, which is the fundamental technique for fitting the SIVC model (1.2) in our thesis. Secondly, in order to avoid overfitting and to select the true model in sparse SIVC models, we will introduce the penalised least squares approach with smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the relevant literature is reviewed in Section 2.2. Thirdly, since determining how to select the tuning parameters (regularization parameters) involved in the SCAD penalty is essential to consistently identify the true model, we refer to Fan and Tang (2013) for the study on generalised information criterion (GIC). A brief review on GIC shall be given in Section 2.3. In the end, we will review

some existing work on the statistical methods with varying coefficient models in Section 2.4.

2.1 Framework of local polynomial modelling

In this section, we will review the framework of local polynomial modelling. Belonging to the family of nonparametric modelling, local polynomial regression does not assume a certain functional form of a regression problem. Instead, the regression functions are left unspecified and determined by data. This approach can be successfully applied to describe an unknown function, which could assess whether a parametric method is appropriate or not. This technique is such a useful tool that it can be applied in broad aspects, which include, among others, non-linear time series, generalised linear models, quantile regression and generalised partially linear single-index models.

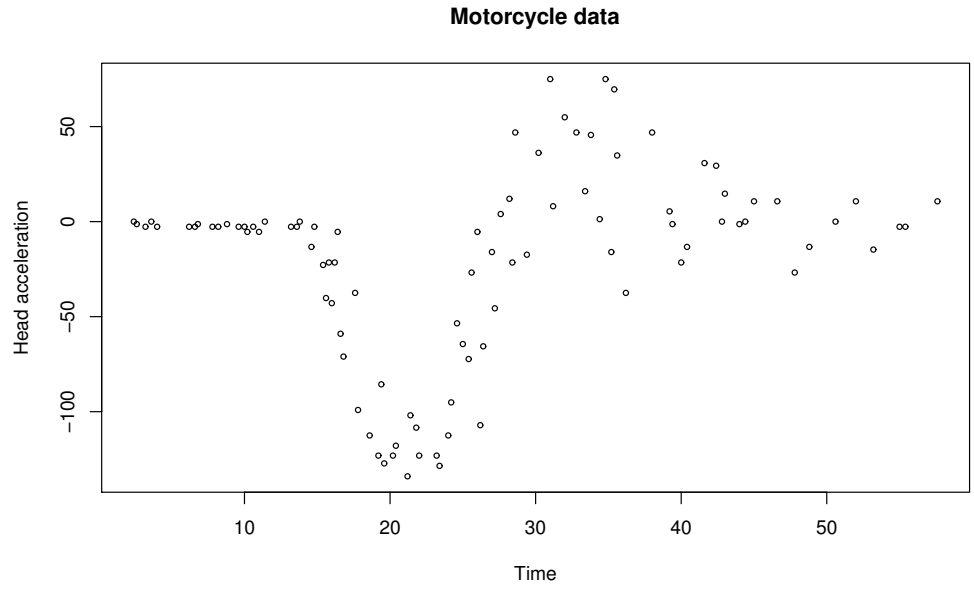
Before outlining the local polynomial regression, we first introduce a motivating example concerning a motorcycle data from Schmidt *et al.*(1981). Two variables are contained in the dataset: the time (in milliseconds) after a simulated impact and the head acceleration, serve as covariate X and response Y, respectively. Figure 2.1 gives the scatter plot diagram of

this dataset. We intuitively find that the observations appear nonlinear, but to gain more insights, we would like to initially fit the data by a linear regression. Assume that (x_i, y_i), $i = 1, \ldots, n$, is the observation collected from the i-th subject. We fit the data by a global linear regression

$$y_i = \alpha_0 + \alpha_1 x_i + \text{error},$$

and report the resulting estimator in Figure 2.2. As illustrated in the first plot of Figure 2.2, the global linear estimates yields a very large modelling bias.

Figure 2.1: Scatter plot for motor data



A commonly used approach to fit the nonlinear phenomena is the polynomial regression. We consider some examples of

polynomial fits as follows:

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \boldsymbol{\epsilon}_i, \tag{2.1}$$

$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3 + \boldsymbol{\epsilon}_i, \tag{2.2}$$

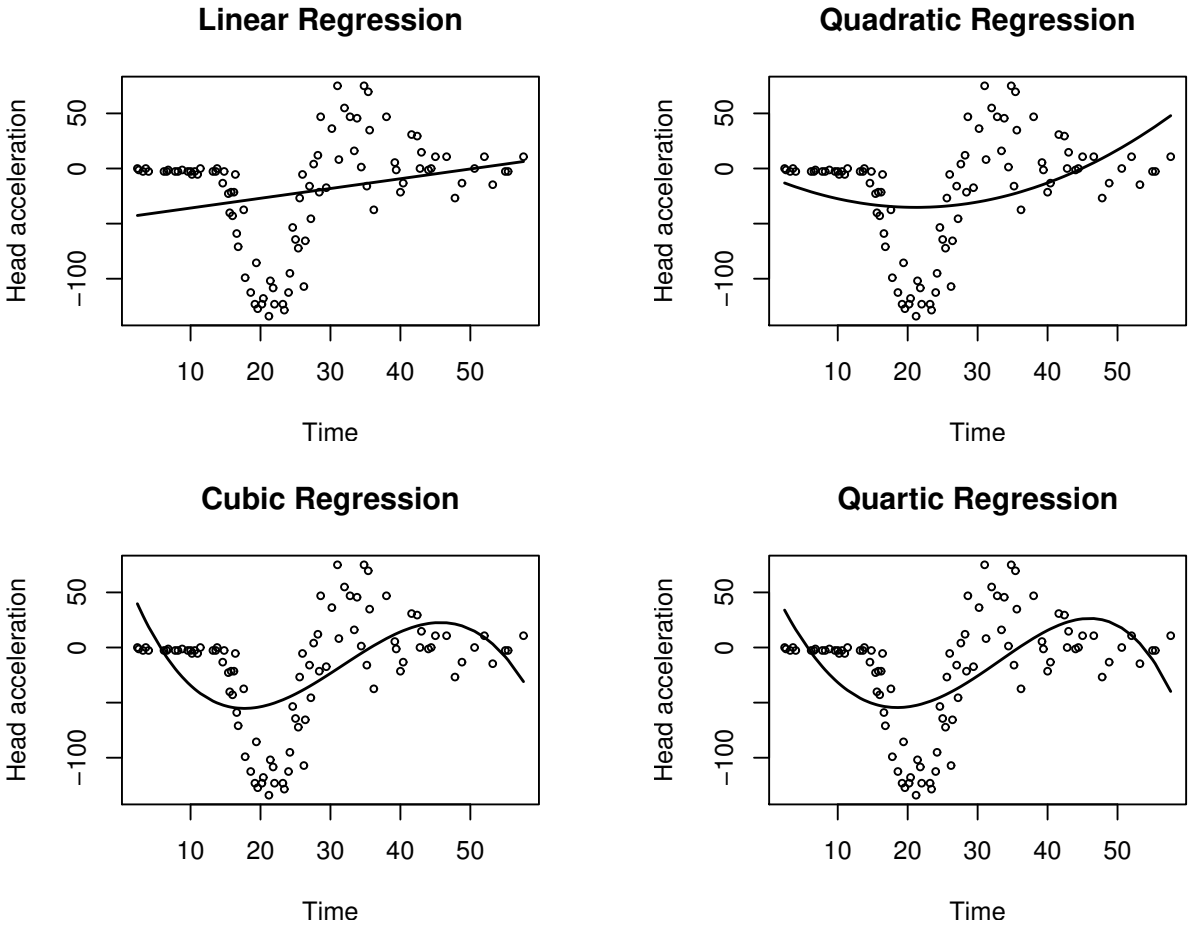
$$y_i = \alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \alpha_3 x_i^3 + \alpha_4 x_i^4 + \boldsymbol{\epsilon}_i, \tag{2.3}$$

where $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_n$ are independently and identically distributed $N(0, \sigma^2)$ random errors. (2.1), (2.2) and (2.3) refers to quadratic, cubic and quartic polynomial regressions respectively. Figure 2.2 shows the estimated curves from them. It can be seen visually that, compared with linear regression, the quadratic, cubic or quartic fit may reduce the modelling bias to some extent, but leads to an estimator with larger variance. Besides, the polynomial models also suffer from the drawback that the remote individual observations can impact largely on the curve.

There are several approaches to overcome the issues of polynomial models. One idea is to apply polynomial model locally to a strip of data around the point that needs to be estimated. We term this method the local (polynomial) modelling. One of the most important hyper-parameter in this modelling is the size of the local neighbourhood, which is called the bandwidth.

To provide more insights into this technique, we apply the local polynomial approximation to an independently and identically distributed bivariate samples $(X_1, Y_1), \cdots, (X_n, Y_n)$

Figure 2.2: Motorcycle data fitted by polynomial regressions



form a population (X, Y). Assume that the data is generated from the model

$$Y = m(X) + \sigma(X)\epsilon, \tag{2.4}$$

where $\mathbb{E}(\epsilon) = 0, \mathrm{Var}(\epsilon) = 1$, and ϵ is independent of X. We wish to fit the unknown regression function $m(x_0) = \mathbb{E}(Y|X = x_0)$ and its derivatives $\dot{m}(x_0), \ddot{m}(x_0), \cdots, m^{(p)}(x_0)$. Suppose that the $(p+1)$th derivative of $m(\cdot)$ exists at the point x_0. Consider a Taylor expansion for the unknown function $m(x)$ for x in a

neighbourhood of x_0

$$m(x) \approx m(x_0) + \dot{m}(x_0)(x - x_0) + \frac{\ddot{m}(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \tag{2.5}$$

We can treat $m(x_0), \dot{m}(x_0), \cdots, m^{(p)}(x_0)$ as unknown parameters that need to be estimated. From this point of view, we use the notation:

$$\frac{m^{(j)}(x_0)}{j!} = \beta_j, \quad \text{for } j = 0, 1, \cdots, p,$$

which allows us to rewrite (2.5) as

$$m(x) \approx \beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \cdots + \beta_p(x - x_0)^p. \tag{2.6}$$

To obtain the estimators of unknown parameters, denoted by $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p$, it suggests minimising a locally weighted least squares regression

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0), \tag{2.7}$$

with respect to β_j, $j = 0, \ldots, p$, where h is a bandwidth, and $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function (a symmetric probability density function) assigning weights to each observation. Based on the estimates $\hat{\beta}_j$, we can obtain the estimator of function $m(x)$ and its derivatives $m^{(v)}(x_0)$ by $\hat{m}_v(x_0) = v!\hat{\beta}_v$ for each $v = 0, \cdots, p$. Using the notations in Fan and Gijbels (1996), the

weighted least squares problem (2.7) can be rewritten in the matrix notation as

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix},$$

$$\mathbf{y} = (Y_1, \cdots, Y_n)^\top,$$
$$\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^\top,$$

and

$$\mathbf{W} = \mathrm{diag}\{K_h(X_1 - x_0), \ldots, K_h(X_n - x_0)\}.$$

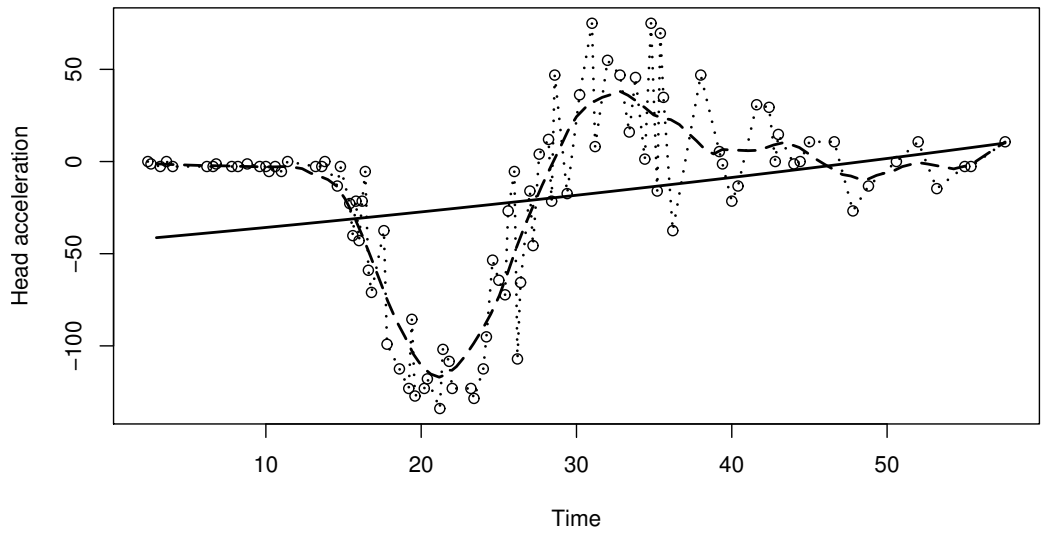
It follows from least squares theory that the solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}. \tag{2.8}$$

To consistently and effectively fit the data by local polynomial regression, it is necessary to choose an appropriate bandwidth h, because it controls the model complexity. A small bandwidth leads to low bias but high variance. A large bandwidth gains on variance side but loses on bias side. Intuitively, we search for a bandwidth which can provide a good

trade-off between bias and variance. Figure 2.3 illustrate this statement by applying the local linear model to the motorcycle data for a variety of bandwidths. We can see from Figure 2.3 that when a very large bandwidth is used, the fit almost yields global linear estimates. Conversely, once $h = 0$ was used, the estimator exactly interpolates the data points. When the bandwidth is chosen to be $h = 3.3$, the local linear regression gives a much more accurate fitting, and hence produces a much smaller approximation error.

Figure 2.3: Local linear regression with different bandwidths



NOTE: *Local linear estimates with bandwidth* $h = 0$ *(dotted line),* 3.3 *(dashed line) and infinity (solid line). With the increasing of h, the estimated curve becomes simpler*

A theoretical optimal choice of bandwidth is obtained

by minimizing the conditional Mean Squared Error (MSE), which is the sum of conditional bias and conditional variance

$$[\text{Bias}(\hat{m}_v(x_0)|\mathbb{X})]^2 + \text{Var}(\hat{m}_v(x_0)|\mathbb{X}),$$

where $\mathbb{X} = (X_1, \ldots, X_n)$. In the practical implementation, bandwidth can be selected by cross validation or generalised cross validation (GCV), and it may be sufficient for some purposes to choose h to be around 25% of the whole range of data.

To deal with the problem of bandwidth selection, it is of importance to have a good insight into bias and variance. The conditional bias and variance of $\hat{\beta}$ can be obtained from (2.8) that

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\beta}}|\mathbb{X}) &= (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{W} m \\ &= \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{r} \end{aligned} \tag{2.9}$$

And

$$\text{Var}(\hat{\boldsymbol{\beta}}|\mathbb{X}) = (\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1}(\mathbf{X}^\top \Sigma \mathbf{X})(\mathbf{X}^\top \mathbf{W}\mathbf{X})^{-1} \tag{2.10}$$

where $m = \{m(\mathbf{X}_1), \cdots, m(\mathbf{X}_n)\}^\top$, $\boldsymbol{\beta} = \{m(x_0), \cdots, m^{(p)}(x_0)/p!\}^\top$, $\mathbf{r} = \mathbf{m} - \mathbf{X}\boldsymbol{\beta}$, the vector of residuals of the local polynomial regression, and $\Sigma = \text{diag}\{K_h^2(X_1 - x_0)\sigma^2(X_1), \ldots, K_h^2(X_n - x_0)\sigma^2(X_n)\}$. Due to the unknown quantities $\mathbf{r}$ and Σ, the expression (2.9) and (2.10) cannot be directly used. The study from Ruppert

and Wand (1994) provides a solution to use a first order asymptotic expansion for the bias and variance to approximate the conditional bias and variance, which is given in the following theorem. The theorem is directly quoted from Fan and Gijbels (1996). We use the following notation:

$$u_j = \int u^j K(u)du, \quad v_j = \int u^j K^2(u)du, \quad S = (u_{j+l})_{0\le j,l\le p},$$

$$\begin{aligned}
&\tilde{S} = (u_{j+l+1})_{0\le j,l\le p}, \quad S^* = (u_{j+l+1})_{0\le j,l\le p} bgv,\\
&c_p = (\mu_{p+1},\cdots,\mu_{2p+1})^\top, \quad \tilde{c}_p = (\mu_{p+2},\cdots,\mu_{2p+2})^\top,\\
&e_{v+1} = (0,\cdots,0,1,0,\cdots,0)^\top,
\end{aligned}$$

where e_{v+1} has a 1 on the $(v+1)^th$ position. $o_p(1)$ denotes a random quantity that is tending to zero in probability.

Theorem 1. *Assume that $f(x_0) > 0$ and that $f(\cdot), m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighbourhood of x_0. Further assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\hat{m}_v(x_0)$ is given by*

$$\begin{aligned}
Var(\hat{m}_v(x_0)|\mathbb{X}) = e_{v+1}^\top S^{-1} S^* S^{-1} e_{v+1} \frac{v!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2v}}\\
+ o_p\left(\frac{1}{nh^{1+2v}}\right). \qquad (2.11)
\end{aligned}$$

The asymptotic conditional bias for $p-v$ odd is given by

$$Bias(\hat{m}_v(x_0)|\mathbb{X}) = e_{v+1}^{\top} S^{-1} c_p \frac{v!}{(p+1)!} m^{(p+1)}(x_0) h^{p+1-v} + o_p(h^{p+1-v}). \tag{2.12}$$

Further, for $p-v$ even the asymptotic conditional bias is

$$Bias(\hat{m}_v(x_0)|\mathbb{X}) = e_{v+1}^{\top} S^{-1} \tilde{c}_p \frac{v!}{(p+1)!} \{m^{(p+2)}(x_0) + (p+2) m^{(p+1)}(x_0) \frac{\dot{f}(x_0)}{f(x_0)}\} h^{p+2-v} + o_p(h^{p+2-v}) \tag{2.13}$$

provided that $\dot{f}(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighbourhood of x_0 and $nh^3 \to \infty$.

We can find from the above theorem that there is a theoretical distinction between the cases $p-v$ odd and the $p-v$ even. Indeed, it turns out later that odd order fits are always superior to even order fits.

2.2 Penalized least squares

The penalised least squares approach is one of the most widely used selection and shrinkage method. This approach attempts to simultaneously select significant variables consistently and estimate the corresponding coefficients effectively. In this

section, we will compactly review the penalised least squares and the smoothly clipped absolute deviation (SCAD) penalty. Start with considering the linear regression model

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ is an $n \times 1$ vector, $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_d)$ is an $n \times d$ design matrix of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top$ is an $d \times 1$ vector of parameters to be estimated and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors.

The penalised least squares (PLS) assumes that

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \Big\{ \frac{1}{2n} ||\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \sum_{k=1}^{d} p_\lambda(|\beta_k|) \Big\}, \tag{2.14}$$

where $p_\lambda(\cdot)$ is a penalty function allowed to depend on the tuning (regularization) parameter $\lambda \geq 0$. The first terms in (2.14) measure the goodness of fit while the second terms control the complexity of the model. Hence, we can regard the minimizer of (2.14) as a trade-off between bias and variance.

To gain the insights about the variable selection procedures more accessible, we consider the specific case of a canonical linear model with a rescaled orthonormal design matrix, i.e., $\mathbf{X}^\top \mathbf{X} = n I_d$. With this in mind, the penalised least squares (2.14) can be rewritten in a minimisation problem as follows:

$$\min_{\boldsymbol{\beta}} \Big\{ \frac{1}{2n} ||\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||_2^2 + \frac{1}{2} ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 + \sum_{k=1}^{d} p_\lambda(|\beta_k|) \Big\}. \tag{2.15}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X^T X})^{-1}\mathbf{X}^\top \mathbf{y} = \mathbf{n}^{-1}\mathbf{X}^\top \mathbf{y}$ is the ordinary least squares estimator. Since (2.15) can be minimised in a component-wise manner, we consider the minimisation problem of the univariate penalise least squares for brevity

$$\frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|), \tag{2.16}$$

with respect to the parameter θ, where z is the univariate ordinary least squares estimate. Then, we can obtain the penalised estimator $\hat{\theta}$ by solving

$$\hat{\theta} = \arg\min_\theta \left\{ \frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \right\}. \tag{2.17}$$

According to the rule provided by Antoniadis and Fan (2001), the penalty function $p_\lambda(\cdot)$ in (2.17) can be clarified as a good penalty function if the corresponding penalised estimate $\hat{\theta}$ can fulfil the following three requirement:

- Sparsity. If the true parameter $|\theta|$ is small, the corresponding resulting estimate will be $\hat{\theta} = 0$.

- Approximate unbiasedness. When the unknown parameter $|\theta|$ is sufficiently large, the resulting estimate gives $\theta = z$ with high probability.

- Continuity. The resulting estimate $\hat{\theta}$ is continuous in data z.

More generally, the sparsity refers to the property that the resulting estimator can automatically shrink the small estimated coefficient to zero and thus reduce model complexity. Approximate unbiasedness is the property that the resulting estimate is nearly unbiased especially when the unknown parameter is large. Continuity represents the property that the resulting estimator is continuous in the data. Fan and Li (2001) also provided some insights on the choice of ideal penalty functions, which included a conclusion that a penalty function holds the sparsity conditions must be singular at the origin.

Continuing on these lines, we can assess some of the most commonly used penalty functions. As a member in the family of L_q penalties, L_0 penalty

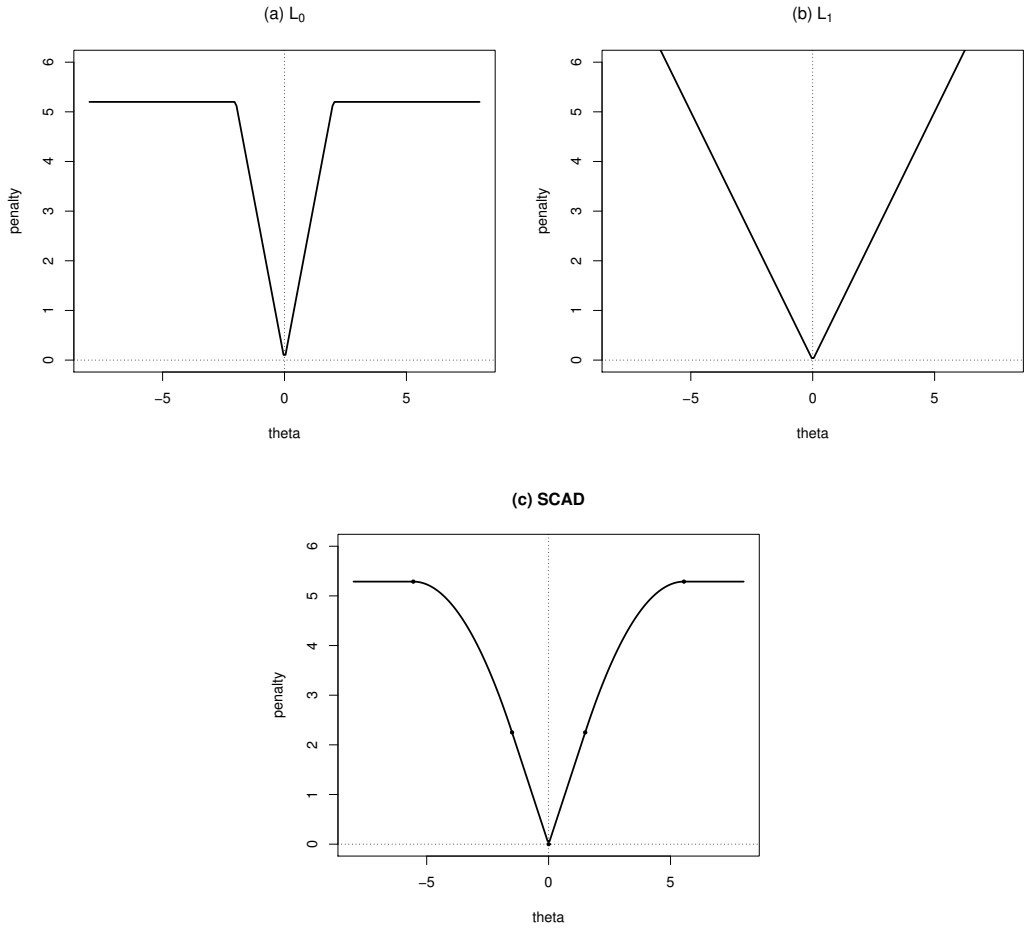
$$p_\lambda(z) = \frac{\lambda^2}{2} I(z \neq 0)$$

produces the hard thresholding estimate $\hat{\theta} = zI(|z| > \lambda)$. Figure 2.4(a) and Figure 2.5(a) visually describes L_0 penalty. It can be see that the resulting estimate does not satisfy the continuity. Another well known penalty is the L_1 penalty (LASSO) (Tibshirani, 1996) $p_\lambda(|\theta|) = \lambda|\theta|$, which yields the soft thresholding estimator

$$\hat{\theta} = \text{sgn}(z)(|z| - \lambda)_+.$$

We depict the thresholding estimate in Figure 2.4(b), from which we can intuitively find that the resulting estimates produce biased solutions. Additionally, the convex L_p penalties with $p > 1$ are not singular around the origin, and hence they fail to enjoy the condition of sparsity. Consequently, None of the L_q penalties can hold all three aforementioned conditions at the same time.

Figure 2.4: The penalty functions



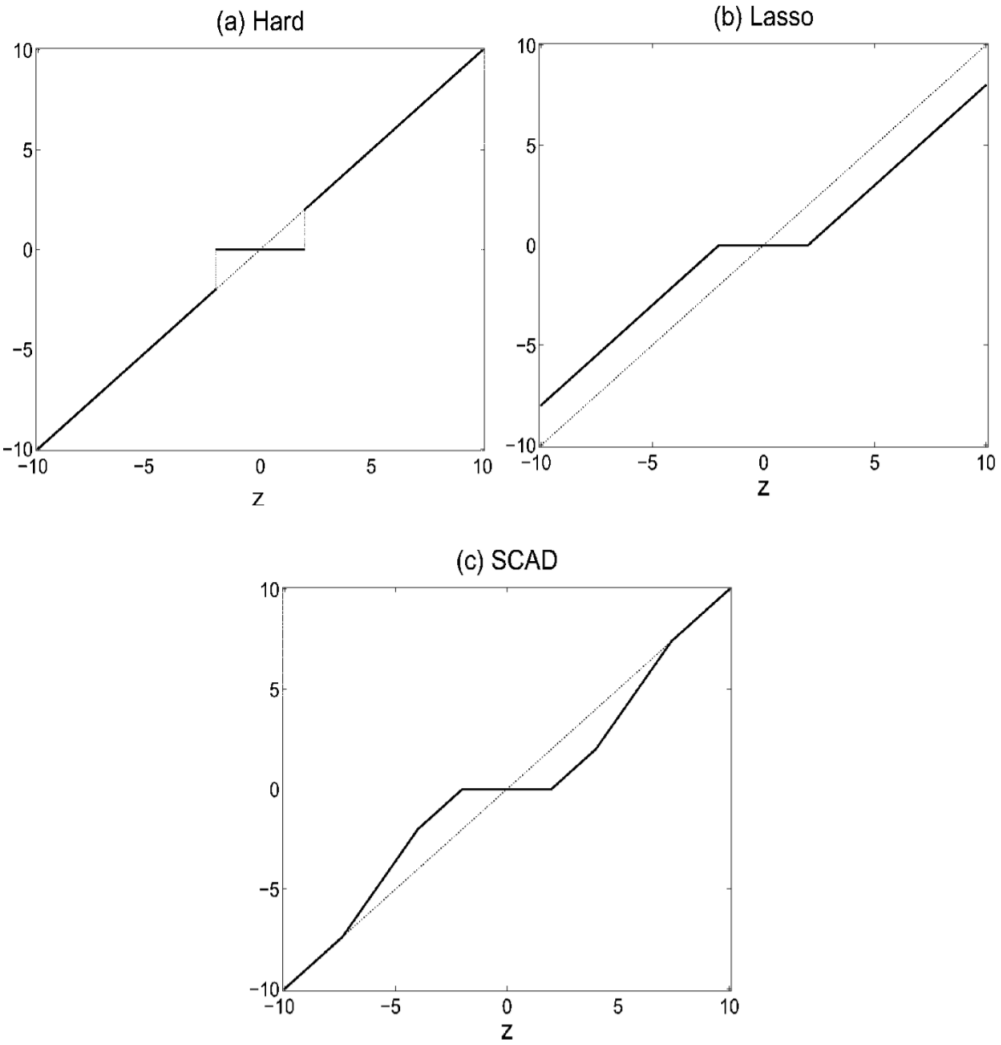
NOTE: *Plot of penalty functions of (a) L_0 penalty, (b) $L_1 penalty$ and (c) $SCAD penalty$.*

As such, one successful attempt, proposed by Fan and Li (2001), is the smoothly clipped absolute deviation (SCAD) penalty, whose derivative is defined by

$$p'_\lambda(\theta) = \lambda\{I(\theta \le \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda)\}, \quad \text{for some } a > 2 \text{ and } \theta > 0,$$

where $p_\lambda(0) = 0$ and a is suggested to be 3.7. It fulfils the foregoing three conditions and, particularly, modifies the bias problems of convex penalties. We gives more insights into this statement by Figure 2.4(c) and 2.5(c).

Figure 2.5: The thresholding functions



NOTE: *Plot of thresholding function for (a) the hard, (b) the soft and (c) the SCAD. The plots are quoted from the Figure 2 in Fan and Li (2001)*

Moreover, Fan and Li (2001) established the asymptotic properties to show that the resulting estimator of SCAD penalty performs as well as the oracle estimator with probability tending to 1. Here, the oracle estimator represents the estimator obtained from the correct sub-model.

Although the SCAD penalty enjoys many appealing properties, solving the penalised least squares (2.14) with a non-convex penalty function is challenging. To solve the minimisation problem, Fan and Li (2001) developed a unified algorithm via local quadratic approximations (LQA).

We assume that a given initial value $\boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_d^0)^\top$ is close to the optimizer of (2.14) and we set $\beta_j = 0$ if β_j^0 is close to 0. Then, the penalty function $p_\lambda(\cdot)$ can be locally approximated by a quadratic function as

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^0|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^0|)}{|\beta_j^0|}[\beta_j^2 - (\beta_j^0)^2], \quad \text{for } \beta_j \approx \beta_j^0. \tag{2.18}$$

The derivative form of this approximation is given as

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\text{sgn}(\beta) \approx \{p'_\lambda(|\beta_j^0|)/|\beta_j^0|\}\beta_j.$$

With this quadratic approximation (2.18), the penalised least squares problem (2.14) is reduced to a quadratic optimisation problem and admits a closed-form solution. Note that one drawback of LQA is that once a coefficient is shrunken to zero in any iteration, it will remain zero. To overcome this potential issue, Zou and Li (2008) developed a unified algorithm based on the local linear approximation (LLA):

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_j^0|) + p'_\lambda(|\beta_j^0|)[|\beta_j| - |\beta_j^0|], \quad \text{for } \beta_j \approx \beta_j^0.$$

It has been demonstrated in Zou and Li (2008) that the LLA does not have to eliminate any small parameters or select the size of perturbation and the LLA naturally yields a sparse estimates through continuous penalisation. Like LQA, the LLA algorithm can also significantly reduce the computation burden.

2.3 Tuning parameter selection by Generalised information criterion

In the previous section, we have discussed penalised least squares with SCAD penalty, which is illustrated to be a remarkably potent shrinkage and selection method. However, many advantages and notable features of the SCAD approach largely depend on a proper choice of the tuning parameters. Traditional model selection criterion includes cross-validation, Akaike information criterion (AIC) (Akaike, 1973) and Bayes information criterion (BIC) (Schwarz, 1978). Wang *et al.*(2007) showed that tuning parameters determined by the BIC could consistently identify the true model for SCAD approach in fixed dimensionality, while AIC and cross-validation may fail because of overfitting. Although a modified BIC still work successfully in diverging dimensionality, when the dimension

of covariates is larger than the sample size, it may fail to select the correct model with consistency and efficiency. To solve this problem, the study of Fan and Tang (2013) allows the dimensionality d increase exponentially with the sample size n and proposed their generalised information criterion (GIC) to select the tuning parameter in high dimensional penalised approach. In Nishii (1984), a generalised information criterion can be expressed as follows:

$$\text{measure of model fitting} + a_n \times \text{measure of model complexity}, \tag{2.19}$$

where a_n is some sequence that controls the regularization on model complexity, and thus the choice of a_n is significant for detecting the optimal tuning parameter. In AIC and BIC, a_n in criterion (2.19) is 2 and $\log(n)$, respectively. Fan and Tang (2013) specified a range of a_n for consistent and effective model selection and proposed a uniform choice

$$a_n = \log\{\log(n)\}\log(d)$$

in GIC (2.19) for practical implementation.

2.4 Varying coefficient models

The varying coefficient model is an important generalisation of the linear model whose coefficients are allowed to be

functions with respect to some random variable U. The nonparametric estimation in varying coefficient model has been well studied in much existing literature. In this section, we provide a concise review of varying coefficient models.

A typical varying coefficient model assumes the following conditional linear structure

$$Y = \sum_{k=1}^{d} f_k(U)X_k + \epsilon, \tag{2.20}$$

for the univariate index variables U, covariates $X_1, \cdots, X_d$ and response variable Y with

$$\mathbb{E}(\epsilon|U, X_1, \cdots, X_d) = 0, \quad \text{Var}(\epsilon|U, X_1, \cdots, X_d) = \sigma^2(U).$$

And we note that it is possible for us to consider an intercept by setting $X_1 \equiv 1$.

Because of the varying coefficient model is equipped with good interpretation, it can be applied to explore the dynamic pattern in many scientific areas where statistics are needed. For instance, in longitudinal data analysis, the coefficient functions $f_k(\cdot)$, $k = 1, \ldots, d$, present the dynamic impact of the corresponding covariate on the response variable over time. When it comes to the estimation of these functional coefficients $f_k(\cdot)$, we can directly fit them by the kernel regression locally around the index U.

Suppose that we have a sample $(U_i, x_{i1}, \ldots, x_{id}, y_i)$, $i = 1, \ldots, n$ from $(U, X_1, \ldots, X_d, Y)$ in model (2.20), then following

the local linear smoothing in Fan and Zhang (1999), for each given u, we locally approximate the function by

$$f_k(U_i) \approx a_k + b_k(U_i - u)$$

for U_i in a neighbourhood of u. This leads to the local estimation procedure with the smooth parameter (bandwidth) h as follows

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{k=1}^{d} [a_k + b_k(U_i - u)] x_{ik} \right\}^2 K_h(U_i - u) \tag{2.21}$$

The locally weighted least squares (2.21) can be rewritten as

$$\min_{\boldsymbol{\theta}} (\mathbf{y} - \mathbb{X}\boldsymbol{\theta})^\top W (\mathbf{y} - \mathbb{X}\boldsymbol{\theta})$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and

$$\begin{aligned} \boldsymbol{\theta} &= (a_1, b_1, \cdots, a_d, b_d)^\top, \\ W &= \text{diag}\{K_h(U_1 - u), \cdots, K_h(U_n - u)\}, \end{aligned}$$

$$\mathbb{X} = \begin{pmatrix} x_{11} & x_{11}(U_1 - u) & \cdots & x_{1d} & x_{1p}(U_1 - u) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n1} & x_{n1}(U_n - u) & \cdots & x_{nd} & x_{np}(U_n - u) \end{pmatrix}.$$

The solution is given by the least squares theory that

$$\hat{\boldsymbol{\theta}} = (\mathbb{X}^\top W \mathbb{X})^{-1} \mathbb{X}^\top W \mathbf{y}$$

and the estimate of coefficient function $f_k(u)$ is

$$\hat{f}_k(u) = e_{2k-1,2d}^{\top}(\mathbb{X}^{\top}W\mathbb{X})^{-1}\mathbb{X}^{\top}W\mathbf{y} \tag{2.22}$$

where $e_{j,m}$ is the unit vector of length m with the $j-$th component being 1.

In traditional varying coefficient models, the index variable U is given to be known. For the purpose of ameliorating the "curse of dimensionality", we introduce the single index model (Hardle and Stoker, 1990) to incorporate with the varying coefficient models. The single index models can be expressed by the following basic form

$$Y = f(X^{\top}\boldsymbol{\beta}_1, \ldots, X^{\top}\boldsymbol{\beta}_q, \epsilon),$$

where X is a d dimensional covariate, Y is the response variable, q is an integer smaller than the dimension d and ϵ is the random error. Hence, the known index U is replaced by the linear combination of covariates and index direction $\boldsymbol{\beta}$, which takes the form $\boldsymbol{\beta}^{\top}\mathbf{X}$. By assuming the index coefficient $\boldsymbol{\beta}$ is unknown and estimated by data, Fan *et al.*(2003) explored the adaptive varying coefficient model (or single index varying coefficient model).

Specifically, suppose that we are going to estimate a multivariate regression function $G(x) \equiv \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$, where Y is a random variable and $\mathbf{X}$ is a $d \times 1$ random vector. The adaptive varying coefficient linear model in Fan *et al.*(2003) which can be one way to approximate $G(\mathbf{x})$ follows the model structure

$$g(\mathbf{x}) = \sum_{k=0}^{d} f_k(\boldsymbol{\beta}^\top \mathbf{x}) x_k \tag{2.23}$$

where $\mathbf{x} = (x_1 \cdots x_d)^\top, x_0 = 1$, $\boldsymbol{\beta} \in \mathbb{R}^d$ is the vector of unknown index parameters and coefficients $f_0(\cdot), \cdots, f_d(\cdot)$ are unknown functions. We obtain the estimators of coefficient functions $f_k(\cdot)$ and index parameters $\boldsymbol{\beta}$ such that $\mathbb{E}\{G(\mathbf{X}) - g(\mathbf{X})\}^2$ is minimised. we remark that once $\boldsymbol{\beta}$ has successfully been fitted, model (2.23) becomes a varying coefficient model (2.20) which can be estimated via the aforementioned local liner regression.

A crucial theorem for the identifiability of the functions $f_k(\cdot)$ are developed in Fan *et al.*(2003). We quote this theorem as follows:

Theorem 2. *Assume $g(\cdot)$ of the form (2.23) to be twice differentiable, if we set $\|\boldsymbol{\beta}\| = 1$, and the first non-zero component of $\boldsymbol{\beta}_0$ positive, such a β is unique unless $g(\cdot)$ is of the following form*

$$g(\mathbf{x}) = \boldsymbol{\alpha}^\top \mathbf{x} \boldsymbol{\beta}^\top \mathbf{x} + \boldsymbol{\gamma}^\top \mathbf{x} + \mathbf{c}, \tag{2.24}$$

where $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathbb{R}^d$, $c \in \mathbb{R}$ are constants and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not parallel to each other. Moreover, once $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$ has been given with $\beta_d \neq 0$, we may let $f_d(\cdot) \equiv 0$. Accordingly, all the other $f_k(\cdot)$ are uniquely determined.

It follows from the Theorem 2 that, in model (2.23), $\|\boldsymbol{\beta}\| = 1$ and the first non-zero element of $\boldsymbol{\beta}$ is positive. To avoid losing the uniqueness of the index parameter $\boldsymbol{\beta}$, it also assumes that the unique least squares approximation $g(\cdot)$ of $G(\cdot)$ should not be formulated in the form (2.24), and hence by letting $\beta \neq 0$, they only consider an approximation $g(\cdot)$ in the following form:

$$g(\mathbf{x}) = \sum_{k=0}^{d-1} f_k(\boldsymbol{\beta}^\top \mathbf{x}) x_k. \tag{2.25}$$

In Fan *et al.*(2003), they not only search for $f_k(\cdot)$ based on the local linear regression, but also give an estimation procedure for $\boldsymbol{\beta}$. As one product of this thesis, we propose an iterative estimation procedure for fitting the model (2.25).

CHAPTER 3

ESTIMATION FOR SINGLE-INDEX VARYING COEFFICIENT MODELS

In this chapter, we first describe the SIVC model. Then, based on the non-parametric estimation, we develop an iterative algorithm to estimate the model. Exploring this algorithm has threefold purposes. Firstly, it yields an efficient approach to solve the varying coefficient model whose index is unknown. Secondly, it helps us to gain insights into the iterative shrinkage estimation procedures which will be generalized in Chapter 4. Thirdly, in our simulation studies, we also employ this penalty-free iterative procedure to estimate the true submodel directly to obtain the oracle estimates, which will be used as a benchmark to evaluate the estimation accuracy of our proposed penalised approach.

3.1 Model specification

Let Y denote the response variable, and $\mathbf{X} = (X_1, \cdots, X_d)^\top$ be a real-valued $d \times 1$ covariate vector with a compact support $\mathscr{D}$, where $\top$ denotes the transpose of a matrix. In a single-index varying coefficient model (SIVCM), it is assumed that for any $\mathbf{x} = (x_1, \cdots, x_d)^\top \in \mathscr{D}$, the regression equation $g(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ admits the following structure:

$$g(\mathbf{x}) = \sum_{k=0}^{d} f_k(\mathbf{x}^\top \boldsymbol{\beta}_0) x_k, \tag{3.1}$$

where $x_0 \equiv 1$, $f_k(\cdot), k = 0, \cdots, d$, are unknown functions, and $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0d})^\top \in \mathbb{R}^d$ is the unknown index parameter. For identification purposes (Fan et al., 2003), if we choose $\|\boldsymbol{\beta}_0\| = 1$, the first non-zero component of $\boldsymbol{\beta}_0$ positive and give $\beta_{0d} \neq 0$, then by setting $f_d(\cdot) \equiv 0$, we can make sure that $f_k(\cdot),\ k \in S_0 \equiv \{0, 1, \cdots, d-1\}$ are uniquely determined, in which case (3.1) is then rewritten as

$$g(\mathbf{x}) = \sum_{k=0}^{d-1} f_k(\mathbf{x}^\top \boldsymbol{\beta}_0) x_k, \tag{3.2}$$

and without loss of generality, it is assumed that $\|\boldsymbol{\beta}_0\| = 1$, $\beta_{01} > 0$, and $\beta_{0d} \neq 0$. Throughout this thesis, we assume all functions $f_k(\cdot)$ are continuously differentiable. For any $k \in S_0$, denote by $\dot{f}_k(\cdot)$, the first order derivative of $f_k(\cdot)$.

In this chapter, we focuses on the estimation of $\boldsymbol{\beta}_0$ as well as of the functions $f_k(\cdot),\ k \in S_0$.

3.2 Methodology

In this section, we outline the approach for estimating the direction $\boldsymbol{\beta}_0$ and functional coefficients $f_k(\cdot)$ in model (3.2). We remark that once the true value of index parameter $\boldsymbol{\beta}_0$ is known, model (3.2) becomes the a typical varying coefficient model with a known index $z = \mathbf{x}^\top \boldsymbol{\beta}_0$, whose coefficient functions can be estimated via local linear regression. Hence, We will first explore the local linear estimators for $f_k(\cdot)$ with given $\boldsymbol{\beta}_0$ in Section 3.2.2, and then extend the idea to an iterative procedure for fitting $f_k(\cdot)$ when $\boldsymbol{\beta}_0$ is unknown in Section 3.2.2.

3.2.1 Estimators for functional coefficients $f_k(\cdot)$ with known $\boldsymbol{\beta}_0$

Let $(\mathbf{X}_i, Y_i), i = 1,, \ldots n$, be independent identical distributed observations with the same marginal distribution as $(\mathbf{X}, Y)$ with $\mathbf{X}_i = (X_{i1}, \cdots, X_{id})^\top$. If the true value of $\boldsymbol{\beta}_0$ is known, then for any given $\mathbf{x}$, the estimation of $\{f_k(\mathbf{x}^\top \boldsymbol{\beta}_0), \dot{f}_k(\mathbf{x}^\top \boldsymbol{\beta}_0)\}$, $k = 0, \ldots, d-1$, can be based on the following Taylor expansion in a neighborhood of $\mathbf{x}^\top \boldsymbol{\beta}_0$ with

$$f_k(z) \approx f_k(\mathbf{x}^\top \boldsymbol{\beta}_0) + \dot{f}_k(\mathbf{x}^\top \boldsymbol{\beta}_0)(z - \mathbf{x}^\top \boldsymbol{\beta}_0).$$

We define $\mathbf{f} = (f_0(\mathbf{x}^\top \boldsymbol{\beta}_0), \ldots, f_{d-1}(\mathbf{x}^\top \boldsymbol{\beta}_0), \dot{f}_0(\mathbf{x}^\top \boldsymbol{\beta}_0), \ldots, \dot{f}_{d-1}(\mathbf{x}^\top \boldsymbol{\beta}_0))^\top \in \mathbb{R}^{2d}$ and write $\tilde{\mathbf{X}}_i = [\mathbf{X}_{i0}, \mathbf{X}_{i1}, \ldots, \mathbf{X}_{id-1}]^\top \in \mathbb{R}^d$, with $\mathbf{X}_{i0} \equiv 1$. Then, the estimator of $\mathbf{f}$ with the condition that $\boldsymbol{\beta}_0$ is known can be obtained by minimizing

$$Q_{\mathbf{x}}(\mathbf{f} \mid \boldsymbol{\beta}_0) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_{i\mathbf{x}}^\top \mathbf{f}\}^2 K_h(\mathbf{X}_{i\mathbf{x}}^\top \boldsymbol{\beta}_0), \tag{3.3}$$

with respect to $\mathbf{f} \in \mathbb{R}^{2d}$, where $K_h(\cdot) = K(\cdot/h)/h$ is a probability density function with the kernel function $K(\cdot)$, h is a smoothing parameter, such that $h \to 0$ as $n \to \infty$, $\mathbf{X}_{i\mathbf{x}} = \mathbf{X}_i - \mathbf{x}$, and

$$\tilde{\mathbf{X}}_{i\mathbf{x}} = [\tilde{\mathbf{X}}_i^\top, (\mathbf{X}_{i\mathbf{x}}^\top \boldsymbol{\beta}_0) \tilde{\mathbf{X}}_i^\top]^\top \in \mathbb{R}^{2d}.$$

Following this idea, estimators of $\mathbf{f}_j \equiv (f_0(\mathbf{X}_j^\top \boldsymbol{\beta}_0), \ldots, f_{d-1}(\mathbf{X}_j^\top \boldsymbol{\beta}_0), \dot{f}_0(\mathbf{X}_j^\top \boldsymbol{\beta}_0), \ldots, \dot{f}_{d-1}(\mathbf{X}_j^\top \boldsymbol{\beta}_0))^\top \in \mathbb{R}^{2d}$, $j = 1, \ldots, n$ can be obtained by implementing n individual minimization of (3.3) with $\mathbf{x}$ replaced by $\mathbf{X}_j$. Then, we consider the local linear estimates for $\mathbf{f}_j$ with the condition that $\boldsymbol{\beta}_0$ is known. This leads to minimizing the double summation of weighted squares

$$Q(\mathbf{F} \mid \boldsymbol{\beta}_0) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_{ij}^\top \mathbf{f}_j\}^2 K_h(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_0), \tag{3.4}$$

with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top \in \mathbb{R}^{n \times 2d}$ with $\mathbf{f}_j = (f_0(\mathbf{X}_j^\top \boldsymbol{\beta}_0), \ldots, f_{d-1}(\mathbf{X}_j^\top \boldsymbol{\beta}_0), \dot{f}_0(\mathbf{X}_j^\top \boldsymbol{\beta}_0), \ldots, \dot{f}_{d-1}(\mathbf{X}_j^\top \boldsymbol{\beta}_0))^\top$ where $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j \in \mathbb{R}^d$ and

$$\tilde{\mathbf{X}}_{ij} = [\tilde{\mathbf{X}}_i^\top, (\mathbf{X}_{ij}^\top \boldsymbol{\beta}_0)\tilde{\mathbf{X}}_i^\top]^\top \in \mathbb{R}^{2d}.$$

We can rewrite (3.4) in matrix notation as

$$\frac{1}{n}\sum_{j=1}^{n}(\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j)^\top W_j(\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j), \tag{3.5}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $\tilde{\mathbb{X}}_j = (\tilde{\mathbf{X}}_{1j}, \ldots, \tilde{\mathbf{X}}_{nj}) \in \mathbb{R}^{2d \times n}$ and $W_j = \text{diag}\{K_h(\mathbf{X}_{1j}^\top \boldsymbol{\beta}_0), \ldots, K_h(\mathbf{X}_{nj}^\top \boldsymbol{\beta}_0)\} \in \mathbb{R}^{n \times n}$. By least squares theory, the resulting estimators $\hat{\mathbf{f}}_j$ is given by

$$\hat{\mathbf{f}}_j = (\tilde{\mathbb{X}}_j W_j \tilde{\mathbb{X}}_j^\top)^{-1}(\tilde{\mathbb{X}}_j W_j \mathbf{Y}),$$

and hence we obtain the estimators $\hat{\mathbf{F}} = (\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_n)^\top$.

With this in mind, it is natural to extend the idea to the case when the index parameter $\boldsymbol{\beta}$ is unknown. However, it's hard to directly work out the estimators for $f_k(\cdot)$ and $\boldsymbol{\beta}$ from an analytic formula, and hence we explore an iterative approach in Section (3.2.2).

3.2.2 Iterative approach for the estimation of SIVC models

In the previous section, we discussed the minimization problem of the locally weighted function (3.4). Now, we consider a similar case but with an unknown index parameter $\boldsymbol{\beta}_0$. We

define the following discrepancy loss function

$$\mathscr{L}(\mathbf{F}, \boldsymbol{\beta}) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_{ij}^{\top} \mathbf{f}_j\}^2 K_h(\mathbf{X}_{ij}^{\top} \boldsymbol{\beta}), \tag{3.6}$$

with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^{\top} \in \mathbb{R}^{n \times 2d}$ and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^{\top} \in \mathbb{R}^d$.

The estimators $\hat{\mathbf{F}}$ and $\hat{\boldsymbol{\beta}}$ are obtained by solving

$$(\hat{\mathbf{F}}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mathbf{F}, \boldsymbol{\beta}} \mathscr{L}(\mathbf{F}, \boldsymbol{\beta}), \tag{3.7}$$

subject to the constraints that $||\boldsymbol{\beta}|| = 1$, $\beta_1 > 0$, which are assumed for the identifiability purpose. A global minimum of the target function (3.6) cannot be derived directly, instead, we consider an iterative computational algorithm to solve the problem. We remark that it is feasible to implement such an iterative approach, as in each step of the iterative procedure, there exists a closed form solution.

Before the iterative procedure, we should specify an initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$. It can be expected that a reasonably good initial value $\tilde{\boldsymbol{\beta}}$ will lead to well performed estimators. We will discuss whether the estimators are sensitive to the choice of initial estimate $\tilde{\boldsymbol{\beta}}$ in the simulation study in Section 7.1.

In order to solve (3.7), we consider the following iterative procedure. Start with an initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$:

1. Step 1: We estimate $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^{\top}$ by solving

$$\hat{\mathbf{F}} = \arg\min_{\mathbf{F}} \mathscr{L}(\mathbf{F}|\tilde{\boldsymbol{\beta}}). \tag{3.8}$$

The estimator denoted by$\hat{\mathbf{F}} = (\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_n)^\top \in \mathbb{R}^{n\times 2d}$ in (3.8) is the minimizer of the following quantity

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_{ij}^\top \mathbf{f}_j\}^2 K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}), \tag{3.9}$$

with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top \in \mathbb{R}^{n\times 2d}$, where $\tilde{\mathbf{X}}_{ij} = [\tilde{\mathbf{X}}_i^\top, (\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}})\tilde{\mathbf{X}}_i^\top]^\top \in \mathbb{R}^{2d}$ and $\tilde{\mathbf{X}}_i = [1, \mathbf{X}_{i1}, \ldots, \mathbf{X}_{i,d-1}]^\top$. The double summation (3.9) can be rewritten in the matrix notation as

$$\frac{1}{n}\sum_{j=1}^{n}(\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j)^\top W_j (\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j),$$

with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top \in \mathbb{R}^{n\times 2d}$ with $\mathbf{f}_j = (f_0(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}), \ldots, f_{d-1}(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}), \dot{f}_0(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}), \ldots, \dot{f}_{d-1}(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}))^\top$, where W_j is an $n \times n$ diagonal matrix with $K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}})$ as its i-th diagonal element, $\tilde{\mathbb{X}}_j$ is an $2d \times n$ matrix with $\tilde{\mathbf{X}}_{ij}$ as its i-th column and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. It follows from the least squares theory that,

$$\hat{\mathbf{f}}_j = (\tilde{\mathbb{X}}_j W_j \tilde{\mathbb{X}}_j^\top)^{-1}(\tilde{\mathbb{X}}_j W_j \mathbf{Y}),$$

and thus we obtain the estimator $\hat{\mathbf{F}} = (\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_n)^\top$. Before next step, we define two $d \times 1$ vectors $\hat{\boldsymbol{a}}_j$ and $\hat{\boldsymbol{b}}_j$ as $\hat{\boldsymbol{a}}_j = (\hat{f}_0(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}), \ldots, \hat{f}_{d-1}(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}))^\top$ and $\hat{\boldsymbol{b}}_j = (\hat{\dot{f}}_0(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}), \ldots, \hat{\dot{f}}_{d-1}(\mathbf{X}_j^\top \tilde{\boldsymbol{\beta}}))^\top$, thereby $\hat{\mathbf{f}}_j = (\hat{\boldsymbol{a}}_j^\top, \hat{\boldsymbol{b}}_j^\top)^\top$.

2. Step 2: By applying the estimator $\hat{\mathbf{F}} = (\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_n)^\top$ with $\hat{\mathbf{f}}_j = (\hat{\boldsymbol{a}}_j^\top, \hat{\boldsymbol{b}}_j^\top)^\top$, $j = 1, \ldots, n$, from Step 1, we now search for the estimator for $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$, by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \mathscr{L}(\boldsymbol{\beta}|\hat{\mathbf{F}}), \tag{3.10}$$

the estimator $\hat{\boldsymbol{\beta}}$ is the minimizer of the function

$$\mathscr{L}(\boldsymbol{\beta}|\hat{\mathbf{F}}) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{a}}_j - \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{b}}_j \mathbf{X}_{ij}^\top \boldsymbol{\beta}\}^2 K_h(\mathbf{X}_{ij}^\top \boldsymbol{\beta}). \tag{3.11}$$

with respect to $\boldsymbol{\beta} = (\beta, \ldots, \beta)^\top \in \mathbb{R}^d$. It worth noting that $\boldsymbol{\beta}$ not only appears in the in the least squares part of the target function (3.11), but also involves in the kernel function. Therefore, it is hard to directly derive a closed form from the quantity (3.11). To deal with this situation, we consider the following approximation for the double summation in (3.11) as

$$\frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{a}}_j - \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{b}}_j \mathbf{X}_{ij}^\top \boldsymbol{\beta}\}^2 K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}), \tag{3.12}$$

with respect to $\boldsymbol{\beta}$. It can be seen that in the kernel function of quantity (3.12), we use the estimate $\tilde{\boldsymbol{\beta}}$ from the last step to replace the unknown parameter $\boldsymbol{\beta}$. Then, $\boldsymbol{\beta}$ only involves in the least squares part, and hence it is feasible to obtain a closed form solution. Unlike the local estimator $\hat{\mathbf{F}}$ which is obtained via the local linear

smoothing, the parameter $\boldsymbol{\beta}$ should be estimated globally. We first rewrite the minimisation problem as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{c_{ij} - \mathbf{B}_{ij}\boldsymbol{\beta}\}^2 w_{ij}, \tag{3.13}$$

where

$$\begin{aligned} c_{ij} &= Y_i - \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{a}}_j, \\ \mathbf{B}_{ij} &= \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{b}}_j \mathbf{X}_{ij}^\top = \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{b}}_j (\mathbf{X}_i - \mathbf{X}_j)^\top \in \mathbb{R}^d, \end{aligned}$$

and

$$w_{ij} = K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) = K_h((\mathbf{X}_i - \mathbf{X}_j)^\top \tilde{\boldsymbol{\beta}}).$$

In order to formulate the double sum of weighted squares in (3.13) into a traditional weighted least squares, we construct an $n^2 \times 1$ vector $\mathbf{C}$, an $n^2 \times d$ matrix $\mathbb{B}$ and an $n^2 \times n^2$ diagonal matrix $\mathbf{W}$ as follows:

$$\begin{aligned} \mathbf{C} &= (c_{11}, \ldots, c_{n1}, c_{12}, \ldots, c_{n2}, \ldots, c_{1n}, \ldots, c_{nn})^\top, \\ \mathbb{B} &= (\mathbf{B}_{11}, \ldots, \mathbf{B}_{n1}, \mathbf{B}_{12}, \ldots, \mathbf{B}_{n2}, \ldots, \mathbf{B}_{1n}, \ldots, \mathbf{B}_{nn})^\top, \\ \mathbf{W} &= \operatorname{diag}\{w_{11}, \ldots, w_{n1}, w_{12}, \ldots, w_{n2}, \ldots, w_{1n}, \ldots, w_{nn}\}. \end{aligned}$$

Then, it leads to a minimisation problem of a traditional weighted least squares as follows:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\mathbf{C} - \mathbb{B}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{C} - \mathbb{B}\boldsymbol{\beta}).$$

Hence, following from least squares theory, the solution is given by

$$\hat{\boldsymbol{\beta}} = (\mathbb{B}^\top \mathbf{W} \mathbb{B})^{-1} (\mathbb{B}^\top \mathbf{W} \mathbf{C}).$$

According to identifiability condition assumed in (3.7), the estimator $\hat{\boldsymbol{\beta}}$ should be rescaled to satisfy the constraints $||\boldsymbol{\beta}|| = 1$, $\beta_1 > 0$.

Then, we go back to Step 1 and replace the estimate $\tilde{\boldsymbol{\beta}}$ with the scaled $\hat{\boldsymbol{\beta}}$ and repeat the two steps until convergence.

CHAPTER 4

MODEL SELECTION IN HIGH-DIMENSIONAL SIVC MODELS

In SIVC models, when the dimension of the covariates is fixed and limited, we can obtain the resulting nonparametric estimators through local linear smoothing as we discussed in the previous chapter. However, if the covariates is of large dimension, because the number of the unknown nonparametric components involved may be exceedingly larger than the number of observations, a direct use of nonparametric modelling may lead to unsatisfactory estimation results. To address this issue, we next introduce a locally weighted group selection method by adding the SCAD penalty to the previous iterative approach to select an efficient predictive model, and thereby to obtain the resulting estimators.

In this chapter, we focus on the main subject of the thesis:

the model selection in high-dimensional SIVC models. Specifically, our model selection includes three aspects: (i) variable selection; (ii) identification of the constant coefficients; (iii) specification of the index.

It is worth noting that variable selection and identification of the constant coefficients are equivalent to detecting the zero functional coefficients and the functional coefficients with zero derivatives respectively. Specification of the index is equivalent to identifying the zero-elements of index parameter $\boldsymbol{\beta}$. In Section 4.1, we give the description of high-dimensional sparse SIVC models. In Section 4.2, we demonstrate the methodology of an iterative computational algorithm for simultaneously selecting and estimating the SIVC model.

4.1 Model specification

Suppose that Y is the response variable, and $\mathbf{X} = (X_1, \cdots, X_{d_n})^\top$ is an $d_n \times 1$ covariate vector. We assume that $d_n \to \infty$ as $n \to \infty$ and d_n is of order $O(n^\alpha)$ for some $0 < \alpha < 1$. For any $\mathbf{x} = (x_1, \cdots, x_{d_n})^\top$, the regression equation $g(\mathbf{x}) \equiv E(Y|\mathbf{X} = \mathbf{x})$ admits the following structure:

$$g(\mathbf{x}) = \sum_{k=0}^{d_n} f_k(\mathbf{x}^\top \boldsymbol{\beta}) x_k, \tag{4.1}$$

where $x_0 \equiv 1$, $f_k(\cdot)$, $k = 0, \cdots, d_n$, are unknown functions, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{d_n})^\top \in \mathbb{R}^{d_n}$ is an unknown vector of index parameters. Let $\beta_{d_n} \neq 0$, it follows the identification condition in model (3.2), to uniquely determine $f_k(\cdot)$, $k \in S_0 \equiv \{0, 1, \cdots, d_n - 1\}$, we rewrite (4.1) as

$$g(\mathbf{x}) = \sum_{k=0}^{d_n - 1} f_k(\mathbf{x}^\top \boldsymbol{\beta}) x_k, \tag{4.2}$$

with $\beta_1 > 0$ and $\|\boldsymbol{\beta}\| = 1$.

It is also assumed that the model (4.2) is a sparse high-dimensional model and, ideally, only a handful predictors contribute to the response or to the index. Therefore, without loss of generality, we assume that there exists a positive integer d_0 which is smaller than d_n, and two subsets S_1 and S_2 of S_0, such that

$$\begin{aligned} &\beta_k \neq 0,\ k = 1, \cdots, d_0, d_n; \quad \beta_k = 0,\ k = d_0 + 1, \cdots, d_n - 1; \\ &S_1 = \{k : k \in S_0,\ f_k(\cdot) \text{ is not constant}\}; \\ &S_2 = \{k : k \in S_0,\ f_k(\cdot) \equiv c_k, \text{ for some } c_k \neq 0\}. \end{aligned} \tag{4.3}$$

For any $k \in S_0$, denote by $\dot{f}_k(.)$, the first order derivative of $f_k(.)$, and let

$$m_k = E[|f_k(\mathbf{X}^\top \boldsymbol{\beta})|],\ \dot{m}_k = E[|\dot{f}_k(\mathbf{X}^\top \boldsymbol{\beta})|],$$

where the expectation is taken with respect to $\mathbf{X}$. Consequently, $k \in S_1 \Leftrightarrow \{m_k > 0, \dot{m}_k > 0\}$, $k \in S_2 \Leftrightarrow \{m_k > 0, \dot{m}_k = 0\}$.

In this chapter, we shall focus on identifying which elements of $\boldsymbol{\beta}$ are zero, and also, among $f_k(\cdot)s$, $k \in S_0$, which are in fact constants or even zero and giving accurate estimates for those nonzero parameters.

4.2 Methodology

Here we introduce the model selection and estimation for SIVC models. The procedure we are going to introduce is a mixture of the ideas of penalised least squares, local linear approximation and group selection. In a similar way to Section 3.2, we initially present our idea on the condition that the true value of index parameter $\boldsymbol{\beta}$ is known, in which the model selection actually becomes selection of the varying-coefficients. By adding appropriate penalty functions to the locally weighted function (3.3), we describe the penalised least squares and obtain the resulting estimators for functional coefficients $f_k(\cdot)$ in Section 4.2.1. Then, we extend the idea to the case that $\boldsymbol{\beta}$ is unknown. In Section 4.2.2, we propose an iterative computational algorithm to simultaneously select and estimate both $f_k(\cdot)$ and $\boldsymbol{\beta}$. At last, in Section 4.2.3, to improve the computational efficiency in the high dimensional situation, we slightly modify the proposed algorithm to reduce its space complexity.

4.2.1 Variable selection and penalised estimators for functional coefficients $f_k(\cdot)$ with known $\boldsymbol{\beta}$

Let $(\mathbf{X}_i, Y_i), i = 1,, \ldots n$, be independent identical distributed observations with the same marginal distribution as $(\mathbf{X}, Y)$ with $\mathbf{X}_i = (X_{i1}, \cdots, X_{id_n})^\top$. For any observation $\mathbf{X}_j$, $j = 1, \ldots, n$, with the true β given, the estimation of $\{f_k(\mathbf{X}_j^\top \boldsymbol{\beta}), \dot{f}_k(\mathbf{X}_j^\top \boldsymbol{\beta})\}$, $k = 0, \ldots, d_n - 1$, can be based on the following Taylor expansion in a neighbourhood of $\mathbf{X}_j^\top \boldsymbol{\beta}$ with

$$f_k(\mathbf{X}_i^\top \boldsymbol{\beta}) \approx f_k(\mathbf{X}_j^\top \boldsymbol{\beta}) + \dot{f}_k(\mathbf{X}_j^\top \boldsymbol{\beta})(\mathbf{X}_i^\top \boldsymbol{\beta} - \mathbf{X}_j^\top \boldsymbol{\beta}). \tag{4.4}$$

We rewrite the Taylor series (4.4) as:

$$f_k(\mathbf{X}_i^\top \boldsymbol{\beta}) \approx a_{jk} + b_{jk}(\mathbf{X}_i - \mathbf{X}_j)^\top \boldsymbol{\beta},$$

where $a_{jk} = f_k(\mathbf{X}_j^\top \boldsymbol{\beta})$, $b_{jk} = \dot{f}_k(\mathbf{X}_j^\top \boldsymbol{\beta})$, $k = 0, \ldots, d_n - 1$ and two $d_n \times 1$ vectors $\boldsymbol{a}_j$ and $\boldsymbol{b}_j$ are denoted as

$$\boldsymbol{a}_j = (a_{j0}, \ldots, a_{j,d_n-1})^\top$$

and

$$\boldsymbol{b}_j = (b_{j0}, \ldots, b_{j,d_n-1})^\top.$$

The local linear estimator of the $2d_n \times 1$ vector

$$\begin{aligned} \mathbf{f}_j &\equiv (f_0(\mathbf{X}_j^\top \boldsymbol{\beta}), \ldots, f_{d_n-1}(\mathbf{X}_j^\top \boldsymbol{\beta}), \dot{f}_0(\mathbf{X}_j^\top \boldsymbol{\beta}), \ldots, \dot{f}_{d_n-1}(\mathbf{X}_j^\top \boldsymbol{\beta}))^\top \\ &= (a_{j0}, \ldots, a_{j,d_n-1}, b_{j0}, \ldots, b_{j,d_n-1})^\top \end{aligned}$$

can be obtained by the minimization of the sum of weighted squares

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{a}_j - (\mathbf{X}_{ij}^\top \boldsymbol{\beta})\tilde{\mathbf{X}}_i^\top \boldsymbol{b}_j\}^2 K_h(\mathbf{X}_{ij}^\top \boldsymbol{\beta}), \tag{4.5}$$

with respect to $\boldsymbol{a}_j, \boldsymbol{b}_j \in \mathbb{R}^{d_n}$, where $K_h(\cdot) = K(\cdot/h)/h$ is a probability density function, h is a smoothing parameter, such that $h \to 0$ as $n \to \infty$, $\mathbf{X}_{ij} = \mathbf{X}_i - \mathbf{X}_j \in \mathbb{R}^{d_n}$ and

$$\tilde{\mathbf{X}}_i = [1, \mathbf{X}_{i1}, \ldots, \mathbf{X}_{i,d_n-1}]^\top \in \mathbb{R}^{d_n}.$$

The problem with this approach, which is inherent to nearly all least square based methods, is that for the zero elements in $\mathbf{f}_j$, their estimates derived from minimizing (4.5) are often not zero. To deal with this issue and produce sparse solutions, so that zero or constant functions could be identified, we combine (4.5) with the smoothly clipped absolute deviation (SCAD) penalty function first proposed in Fan and Li (2001), the derivative of which is such that

$$\dot{p}_\lambda(t) = \lambda\{I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\}, \quad t > 0,$$

for some constant $a > 2$, and a regularization parameter λ; see Fan and Li (2001) for detailed discussions on the various desirable properties of the SCAD penalty function.

This leads to the following locally weighted group - SCAD function for feature selection and identification of constant coefficients

$$Q(\mathbf{F} \mid \boldsymbol{\lambda}, \boldsymbol{\beta}) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{a}_j - (\mathbf{X}_{ij}^\top \boldsymbol{\beta}) \tilde{\mathbf{X}}_i^\top \boldsymbol{b}_j\}^2 K_h(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) + \sum_{k=0}^{d_n-1} p_{\lambda_k}(||\mathbf{a}_{(k)}||) + \sum_{k=0}^{d_n-1} p_{\lambda_{d_n+k}}(||\mathbf{b}_{(k)}||), \tag{4.6}$$

with respect to the $n \times 2d_n$ matrix

$$\mathbf{F} = \begin{pmatrix} \mathbf{a}_1, & \boldsymbol{b}_1 \\ \vdots & \vdots \\ \mathbf{a}_n, & \boldsymbol{b}_n \end{pmatrix} = (\mathbf{a}_{(0)}, \ldots, \mathbf{a}_{(d_n-1)}, \mathbf{b}_{(0)}, \ldots, \mathbf{b}_{(d_n-1)}),$$

where $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_{2d_n-1})^\top$ is the $2d_n \times 1$ vector of regularization (tuning) parameters and $||\cdot||$ stands for the Euclidean norm. We note that the task of selection of varying-coefficient becomes equivalent to detecting spares columns in matrix $(\mathbf{a}_{(0)}, \ldots, \mathbf{a}_{(d_n-1)})$, which is the first d_n columns of matrix $\mathbf{F}$. Direct use of the SCAD method on model (4.2) for general variable selection is not efficient, which leads to select far more individuals than necessary; accordingly, in (4.6), we follow the group selection idea of Yuan and Lin (2006) to identify the sparse solutions in $(\mathbf{a}_{(0)}, \ldots, \mathbf{a}_{(d_n-1)})$ in a column-wise manner. Analogically, we also apply the group selection idea to select $\dot{f}_k(\cdot)$, which is equivalent to identify sparse solution in

matrix $(\mathbf{b}_{(0)}, \ldots, \mathbf{b}_{(d_n-1)})$ in column-wise manner. Consequently, we can obtain the penalised estimate $\hat{\mathbf{F}}$ by solving

$$\hat{\mathbf{F}} = \arg\min_{\mathbf{F}} Q(\mathbf{F} \mid \boldsymbol{\lambda}, \boldsymbol{\beta}). \tag{4.7}$$

In order to deal with the SCAD-type problems, we need to introduce a computational algorithm. For the purpose of simplicity and completeness, we here apply an algorithm based on the idea of the local quadratic approximation proposed by Fan and Li (2001). In (4.6), the penalty function with respect to $\mathbf{a}_{(k)}$ can be locally approximated by a quadratic function as:

$$p_{\lambda_k}(||\mathbf{a}_{(k)}||) \approx p_{\lambda_k}(||\mathbf{a}^0_{(k)}||) + \frac{1}{2}\frac{p'_{\lambda_k}(||\mathbf{a}^0_{(k)}||)}{||\mathbf{a}^0_{(k)}||}[\mathbf{a}^\top_{(k)}\mathbf{a}_{(k)} - (\mathbf{a}^0_{(k)})^\top \mathbf{a}^0_{(k)}], \quad \text{for } \mathbf{a}_{(k)} \approx \mathbf{a}^0_{(k)}, \tag{4.8}$$

where $\mathbf{a}^0_{(k)}$ is an initial value that is supposed to be close to the minimiser $\hat{\mathbf{a}}_{(k)}$ of (4.6). If $\mathbf{a}^0_{(k)}$ is very close to $\mathbf{0}$, we directly set $\hat{\mathbf{a}}_{(k)} = \mathbf{0}$. Alternatively, the local quadratic approximation of the first derivative of $\mathbf{a}_{(k)}$ can be given by

$$[p_{\lambda_k}(||\mathbf{a}_{(k)}||)]' \approx \frac{p'_{\lambda_k}(||\mathbf{a}^0_{(k)}||)}{||\mathbf{a}^0_{(k)}||}||\mathbf{a}_{(k)}||.$$

Similarly, the the penalty function with respect to $\mathbf{b}_{(k)}$ and its

first derivative can be locally approximated by

$$p_{\lambda_{k+d_n}}(||\mathbf{b}_{(k)}||) \approx p_{\lambda_{k+d_n}}(||\mathbf{b}^0_{(k)}||) + \frac{1}{2}\frac{p'_{\lambda_{k+d_n}}(||\mathbf{b}^0_{(k)}||)}{||\mathbf{b}^0_{(k)}||} \times [\mathbf{b}^\top_{(k)}\mathbf{b}_{(k)} - (\mathbf{b}^0_{(k)})^\top \mathbf{b}^0_{(k)}], \quad \text{for } \mathbf{b}_{(k)} \approx \mathbf{b}^0_{(k)}, \tag{4.9}$$

and

$$[p_{\lambda_{k+d_n}}(||\mathbf{b}_{(k)}||)]' \approx \frac{p'_{\lambda_{k+d_n}}(||\mathbf{b}^0_{(k)}||)}{||\mathbf{b}^0_{(k)}||}||\mathbf{b}_{(k)}||,$$

respectively.

Then, by replacing the penalty functions in (4.6) by the approximations (4.8) and (4.9), it leads to the following objective function

$$\begin{aligned} Q(\mathbf{F} \mid \boldsymbol{\lambda}, \boldsymbol{\beta}) = {} & \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{a}_j - (\mathbf{X}_{ij}^\top \boldsymbol{\beta})\tilde{\mathbf{X}}_i^\top \boldsymbol{b}_j\}^2 K_h(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) \\ & + \frac{1}{2}\sum_{k=0}^{d_n-1}\frac{p'_{\lambda_k}(||\mathbf{a}^0_{(k)}||)}{||\mathbf{a}^0_{(k)}||}\mathbf{a}^\top_{(k)}\mathbf{a}_{(k)} + \frac{1}{2}\sum_{k=0}^{d_n-1}\frac{p'_{\lambda_{k+d_n}}(||\mathbf{b}^0_{(k)}||)}{||\mathbf{b}^0_{(k)}||}\mathbf{b}^\top_{(k)}\mathbf{b}_{(k)} \\ & + C, \end{aligned} \tag{4.10}$$

where C stands for the constant terms when the initial values $\mathbf{a}^0_{(k)}$ and $\mathbf{b}^0_{(k)}$ are provided.

We denote $\mathbf{f}_j = (\boldsymbol{a}_j^\top, \boldsymbol{b}_j^\top)^\top \in \mathbb{R}^{2d_n}$, in other words, $\{\mathbf{f}_{(0)}, \ldots, \mathbf{f}_{(2d_n-1)}\} = \{\boldsymbol{a}_{(0)}, \ldots, \boldsymbol{a}_{(d_n-1)}, \boldsymbol{b}_{(0)}, \ldots, \boldsymbol{b}_{(d_n-1)}\}$. Then, in matrix notation, the objective function (4.10) is equivalent to

$$Q(\mathbf{F} \mid \boldsymbol{\lambda}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j)^\top W_j (\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j) + \frac{1}{2} \sum_{j=1}^{n} \mathbf{f}_j^\top \Sigma_{\boldsymbol{\lambda}}(\mathbf{F}^0) \mathbf{f}_j + C, \tag{4.11}$$

with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top = (\mathbf{f}_{(0)}, \ldots, \mathbf{f}_{(2d_n-1)}) \in \mathbb{R}^{n \times 2d_n}$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$, $W_j = \text{diag}\{K_h(\mathbf{X}_{1j}^\top \boldsymbol{\beta}_0), \ldots, K_h(\mathbf{X}_{nj}^\top \boldsymbol{\beta}_0)\} \in \mathbb{R}^{n \times n}$, $\mathbf{F}^0 = (\mathbf{f}_{(0)}^0, \ldots, \mathbf{f}_{(2d_n-1)}^0)$ is the initial value of $\mathbf{F}$, $\tilde{\mathbb{X}}_j$ is an $2d_n \times n$ matrix as

$$\tilde{\mathbb{X}}_j = (\tilde{\mathbf{X}}_{1j}, \ldots, \tilde{\mathbf{X}}_{nj}) \text{ with } \tilde{\mathbf{X}}_{ij} = [\tilde{\mathbf{X}}_i^\top, (\mathbf{X}_{ij}^\top \boldsymbol{\beta}_0) \tilde{\mathbf{X}}_i^\top]^\top,$$

and $\Sigma_{\boldsymbol{\lambda}}(\mathbf{F}^0)$ is an $2d_n \times 2d_n$ diagonal matrix as

$$\Sigma_{\boldsymbol{\lambda}}(\mathbf{F}^0) = \text{diag}\{\frac{p'_{\lambda_0}(||\mathbf{f}_{(0)}^0||)}{||\mathbf{f}_{(0)}^0||}, \ldots, \frac{p'_{\lambda_{2d_n-1}}(||\mathbf{f}_{(2d_n-1)}^0||)}{||\mathbf{f}_{(2d_n-1)}^0||}\}.$$

The solution of the penalised least squares (4.11) can be found by computing the ridge regression

$$\hat{\mathbf{f}}_j = \{\tilde{\mathbb{X}}_j W_j \tilde{\mathbb{X}}_j^\top + \frac{n}{2} \Sigma_{\boldsymbol{\lambda}}(\mathbf{F}^0)\}^{-1} \tilde{\mathbb{X}}_j W_j \mathbf{Y},$$

and hence we obtain the resulting estimators $\hat{\mathbf{F}} = (\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_n)^\top$. By applying the proposed penalised approach, the resulting estimators of insignificant coefficients are expected to be shrunk to a very small value. Sequentially, in our implementation,

we introduce an appropriate threshold which is used to automatically reduce the very small estimators to zero, and thus we obtain the final penalised estimator with sparsity. Specifically, assuming that the thresholds $\{\theta_1, \theta_2\}$ are small enough, if $||\hat{\mathbf{a}}_{(k)}||$, $k = 0, \ldots, d_n - 1$, the L_2-norm of penalised estimator $\hat{\mathbf{a}}_{(k)}$, is smaller than θ_1, we set $||\hat{\mathbf{a}}_{(k)}|| = 0$ which is equivalent to shrinking the estimator of k -th coefficient $f_k(\cdot)$ to zero. Analogously, if $||\hat{\mathbf{b}}_{(k)}|| < \theta_2$ but the corresponding $||\hat{\mathbf{a}}_{(k)}|| \geq \theta_2$, we will consider the k -th coefficient $f_k(\cdot)$ as constant, whose resulting estimator can be approximated as

$$\hat{f}_k(\cdot) = \frac{1}{n} \sum_{j=1}^{n} \hat{a}_{jk}.$$

We remark that, in the foregoing algorithm, a reasonably good initial value $\mathbf{F}^0$ is very important for an efficient estimation. Practically, in the iterative procedure described in next section,we solve an ordinary locally weighted least squares (3.4) before conducting the penalised approach, and then use the minimisers of (3.4) as the initial values. From now on, we call the minimisers of ordinary least squares as "preliminary estimators", which serve as the initial values of our penalised approach. With this in mind, we will specify this method in Section 4.2.2. Meanwhile, it is natural to extend the idea to the case when the index parameter $\boldsymbol{\beta}$ is unknown, and hence we are also going to explore an iterative algorithm for selecting and estimating the SIVC models in Section 4.2.2.

4.2.2 Iterative approach for the model selection and estimation of high-dimensional SIVC model

In this section, we generalise our methodology of model selection to the high-dimensional SIVC model (4.2) to select a sub-model of important components, and thereby produce accurate estimation.

Let

$$\mathscr{L}(\mathbf{F},\boldsymbol{\beta}) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{a}_j - (\mathbf{X}_{ij}^\top\boldsymbol{\beta})\tilde{\mathbf{X}}_i^\top \boldsymbol{b}_j\}^2 K_h(\mathbf{X}_{ij}^\top\boldsymbol{\beta}). \quad (4.12)$$

The penalised local weighted least squares for model selection and estimation is given by

$$\begin{aligned} Q(\mathbf{F},\boldsymbol{\beta}\mid\boldsymbol{\lambda}) \;=\; & \mathscr{L}(\mathbf{F},\boldsymbol{\beta}) + \sum_{k=0}^{d_n-1} p_{\lambda_k}(||\mathbf{a}_{(k)}||) \\ & + \sum_{k=0}^{d_n-1} p_{\lambda_{d_n+k}}(||\mathbf{b}_{(k)}||) + \sum_{k=1}^{d_n} p_{\tilde{\lambda}_k}(|\beta_k|), \quad (4.13) \end{aligned}$$

with respect to $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_{d_n})^\top \in \mathbb{R}^{d_n}$ and

$$\mathbf{F} = \begin{pmatrix} \mathbf{a}_1, & \boldsymbol{b}_1 \\ \vdots & \vdots \\ \mathbf{a}_n, & \boldsymbol{b}_n \end{pmatrix} = (\mathbf{a}_{(0)},\ldots,\mathbf{a}_{(d_n-1)},\mathbf{b}_{(0)},\ldots,\mathbf{b}_{(d_n-1)}) \in \mathbb{R}^{n\times 2d};$$

where $\boldsymbol{\lambda} = (\lambda_0,\lambda_1,\ldots,\lambda_{2d_n-1},\tilde{\lambda}_1,\ldots,\tilde{\lambda}_{d_n})^\top$ is the augmented vector of the pre-specified regularization (tuning) parameters,

and $\left|\cdot\right|$ stands for the L_1-norm. The L_1-norm penalty induces sparsity in the solution of $\boldsymbol{\beta}$. It is worth noting that in (4.13) we directly apply the SCAD method to select the index parameter $\boldsymbol{\beta}$ individually and apply the group selection method to select the functional coefficients and their derivatives, namely, we identify sparse solutions in $\boldsymbol{\beta}$ in a element-wise manner but in matrix $\mathbf{F}$ in a column-wise manner.

Consequentially, the penalised estimators $\hat{\mathbf{F}}$ and $\hat{\boldsymbol{\beta}}$ can be obtained by solving

$$(\hat{\mathbf{F}}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mathbf{F},\boldsymbol{\beta}} Q(\mathbf{F}, \boldsymbol{\beta} \mid \boldsymbol{\lambda}), \tag{4.14}$$

subject to the constraints that $||\boldsymbol{\beta}|| = 1$, $\beta_1 > 0$. As direct minimization of (4.13) is hard to conduct, we propose an iterative procedure for implementation purpose.

Start with an initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$:

1. Step 1: We first work out preliminary estimators of the functional coefficients using ordinary weighted least squares which will serve as the initial values of the penalised approach. The preliminary estimates, which is denoted by $\hat{\mathbf{F}}^o = (\hat{\mathbf{a}}^o_{(0)}, \ldots, \hat{\mathbf{a}}^o_{(d_n-1)}, \hat{\mathbf{b}}^o_{(0)}, \ldots, \hat{\mathbf{b}}^o_{(d_n-1)}) \in \mathbb{R}^{n\times 2d}$ can be obtained by solving

$$\hat{\mathbf{F}}^O = \arg\min_{\mathbf{F}} \mathscr{L}(\mathbf{F}|\tilde{\boldsymbol{\beta}}). \tag{4.15}$$

This leads to minimizing the double sum

$$\frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i-\tilde{\mathbf{X}}_i^\top\mathbf{a}_j-(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}})\tilde{\mathbf{X}}_i^\top\boldsymbol{b}_j\}^2K_h(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}}), \qquad (4.16)$$

with respect to $\{\mathbf{a}_j\}$ and $\{\boldsymbol{b}_j\}$, $j=1,\ldots,n$. $\mathbf{X}_{ij}=\mathbf{X}_i-\mathbf{X}_j\in\mathbb{R}^{d_n}$ and

$$\tilde{\mathbf{X}}_i=[1,\mathbf{X}_{i1},\ldots,\mathbf{X}_{i,d_n-1}]^\top\in\mathbb{R}^{d_n}.$$

For brevity, we recall the pre-specified notation

$$\begin{aligned}&\mathbf{f}_j\equiv(\boldsymbol{a}_j^\top,\boldsymbol{b}_j^\top)^\top,\\&\tilde{\mathbf{X}}_{ij}=[\tilde{\mathbf{X}}_i^\top,(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}})\tilde{\mathbf{X}}_i^\top]^\top,\\&\tilde{\mathbb{X}}_j=(\tilde{\mathbf{X}}_{1j},\ldots,\tilde{\mathbf{X}}_{nj}).\end{aligned}$$

Then, (4.16) can be rewrite in the matrix notation as

$$\frac{1}{n}\sum_{j=1}^{n}(\mathbf{Y}-\tilde{\mathbb{X}}_j^T\mathbf{f}_j)^TW_j(\mathbf{Y}-\tilde{\mathbb{X}}_j^T\mathbf{f}_j),$$

with respect to $\mathbf{F}=(\mathbf{f}_1,\ldots,\mathbf{f}_n)^\top=(\mathbf{f}_{(0)},\ldots,\mathbf{f}_{(2d_n-1)})\in\mathbb{R}^{n\times 2d_n}$, where W_j is an $n\times n$ diagonal matrix with $K_h(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}})$ as its i-th diagonal element and $\mathbf{Y}=(Y_1,\ldots,Y_n)^\top$. It follows from the least squares theory that,

$$\hat{\mathbf{f}}_j^o=(\tilde{\mathbb{X}}_jW_j\tilde{\mathbb{X}}_j^T)^{-1}(\tilde{\mathbb{X}}_jW_j\mathbf{Y}),$$

and thus we obtain the preliminary estimator $\hat{\mathbf{F}}^o=(\hat{\mathbf{f}}_1^o,\ldots,\hat{\mathbf{f}}_n^o)^\top$.

2. Step 2: Based on the idea of the local quadratic approximation and with the help of the preliminary estimator $\hat{\mathbf{F}}^o$, we solve the minimisation problem of the proposed penalised weighted least squares along with group SCAD penalty, and thus obtain estimators of the varying coefficients, which denoted by $\hat{\mathbf{F}}^p = (\hat{\mathbf{f}}_1^p, \ldots, \hat{\mathbf{f}}_n^p)^\top$. Now, we need to estimate

$$\hat{\mathbf{F}}^p = \arg\min_{\mathbf{F}} Q(\mathbf{F} \mid \boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}}, \hat{\mathbf{F}}^o).$$

Specifically, we consider the following locally weighted group-SCAD function with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top = (\mathbf{f}_{(0)}, \ldots, \mathbf{f}_{(2d_n-1)}) \in \mathbb{R}^{n \times 2d_n}$:

$$\begin{aligned} Q(\mathbf{F} \mid \boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}}) = & \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{a}_j - (\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) \tilde{\mathbf{X}}_i^\top \boldsymbol{b}_j\}^2 K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) \\ & + \sum_{k=0}^{d_n-1} p_{\lambda_k}(||\mathbf{a}_{(k)}||) + \sum_{k=0}^{d_n-1} p_{\lambda_{d_n+k}}(||\mathbf{b}_{(k)}||). \end{aligned} \tag{4.17}$$

By applying the quadratic approximation (4.8) and (4.9) to the penalty functions corresponding to $||\mathbf{a}_{(k)}||$ and $||\mathbf{b}_{(k)}||$, respectively, a new objective function can be defined by

$$Q(\mathbf{F} \mid \boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}}, \hat{\mathbf{F}}^o) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{a}_j - (\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) \tilde{\mathbf{X}}_i^\top \boldsymbol{b}_j\}^2 \times K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) + \frac{1}{2} \sum_{k=0}^{d_n - 1} \frac{p'_{\lambda_k}(||\hat{\mathbf{a}}_{(k)}^o||)}{||\hat{\mathbf{a}}_{(k)}^o||} \mathbf{a}_{(k)}^\top \mathbf{a}_{(k)} + \frac{1}{2} \sum_{k=0}^{d_n - 1} \frac{p'_{\lambda_{k+d_n}}(||\hat{\mathbf{b}}_{(k)}^o||)}{||\hat{\mathbf{b}}_{(k)}^o||} \mathbf{b}_{(k)}^\top \mathbf{b}_{(k)} + C, \tag{4.18}$$

where C stands for the constant terms when the preliminary estimator $\hat{\mathbf{F}}^o = (\hat{\mathbf{f}}_1^o, \ldots, \hat{\mathbf{f}}_n^o)^\top$ are provided. Then, in matrix notation, the minimization problem of (4.18) is equivalent to minimizing

$$\frac{1}{n} \sum_{j=1}^{n} (\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j)^\top W_j (\mathbf{Y} - \tilde{\mathbb{X}}_j^\top \mathbf{f}_j) + \frac{1}{2} \sum_{j=1}^{n} \mathbf{f}_j^\top \Sigma_{\boldsymbol{\lambda}}(\hat{\mathbf{F}}^o) \mathbf{f}_j + C, \tag{4.19}$$

with respect to $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^\top = (\mathbf{f}_{(0)}, \ldots, \mathbf{f}_{(2d_n-1)}) \in \mathbb{R}^{n \times 2d_n}$, where $W_j = \mathrm{diag}\{K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}), \ldots, K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}})\} \in \mathbb{R}^{n \times n}$, $\hat{\mathbf{F}}^o = (\hat{\mathbf{f}}_{(0)}^o, \ldots, \hat{\mathbf{f}}_{(2d_n-1)}^o) \equiv (\hat{\mathbf{a}}_{(0)}^o, \ldots, \hat{\mathbf{a}}_{(d_n-1)}^o, \hat{\mathbf{b}}_{(0)}^o, \ldots, \hat{\mathbf{b}}_{(d_n-1)}^o)$ is the preliminary estimator of $\mathbf{F}$ and $\Sigma_{\boldsymbol{\lambda}}(\hat{\mathbf{F}}^o)$ is an $2d_n \times 2d_n$ diagonal matrix as

$$\Sigma_{\boldsymbol{\lambda}}(\hat{\mathbf{F}}^o) = \mathrm{diag}\left\{ \frac{p'_{\lambda_0}(||\hat{\mathbf{f}}_{(0)}^o||)}{||\hat{\mathbf{f}}_{(0)}^o||}, \ldots, \frac{p'_{\lambda_{2d_n-1}}(||\hat{\mathbf{f}}_{(2d_n-1)}^o||)}{||\mathbf{f}_{(2d_n-1)}^o||} \right\}.$$

The solution of the penalised least squares (4.19) can be found by

$$\hat{\mathbf{f}}_j^p = \{\tilde{\mathbb{X}}_j W_j \tilde{\mathbb{X}}_j^T + \frac{n}{2}\Sigma_{\boldsymbol{\lambda}}(\hat{\mathbf{F}}^o)\}^{-1}\tilde{\mathbb{X}}_j W_j \mathbf{Y},$$

and hence we obtain the resulting estimators $\hat{\mathbf{F}}^p = (\hat{\mathbf{f}}_1^p, \ldots, \hat{\mathbf{f}}_n^p)^T$. Note that we have suppressed the dependency of these quantities on $\boldsymbol{\lambda}$.

3. Step 3: By applying the estimator $\hat{\mathbf{F}}^p = (\hat{\mathbf{f}}_1^p, \ldots, \hat{\mathbf{f}}_n^p)^T$ with $\hat{\mathbf{f}}_j \equiv \{(\hat{\mathbf{a}}_j^p)^\top, (\hat{\mathbf{b}}_j^p)^\top\}^\top$, $j = 1, \ldots, n$, we now search for the preliminary estimator for $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}^o$, by solving

$$\hat{\boldsymbol{\beta}}^o = \arg\min_{\boldsymbol{\beta}} \mathscr{L}(\boldsymbol{\beta}|\hat{\mathbf{F}}^p), \tag{4.20}$$

the estimator $\hat{\boldsymbol{\beta}}^o$ is the minimizer of the function

$$\mathscr{L}(\boldsymbol{\beta}|\hat{\mathbf{F}}^p) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{a}}_j^p - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{b}}_j^p \mathbf{X}_{ij}^\top \boldsymbol{\beta}\}^2 K_h(\mathbf{X}_{ij}^\top \boldsymbol{\beta}). \tag{4.21}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_n})^\top \in \mathbb{R}^{d_n}$. Note that $\boldsymbol{\beta}$ not only appears in the in the least squares part of the target function (4.21), but also involves in the kernel function. In order to obtain a closed form solution, we follow exactly the same way in section 3 by approximating the target function (4.21) by

$$\mathscr{L}(\boldsymbol{\beta}|\hat{\mathbf{F}}^p) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{a}}_j^p - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{b}}_j^p \mathbf{X}_{ij}^\top \boldsymbol{\beta}\}^2 K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}), \tag{4.22}$$

where $\tilde{\boldsymbol{\beta}}$ is the estimator for $\boldsymbol{\beta}$ used in Step 1 and Step 2. Rewriting the minimization problem yields

$$\hat{\boldsymbol{\beta}}^o = \arg\min_{\boldsymbol{\beta}} \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{C_{ij} - \mathbf{B}_{ij}\boldsymbol{\beta}\}^2 W_{ij}, \tag{4.23}$$

where

$$\begin{aligned} C_{ij} &= Y_i - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{a}}_j^p, \\ \mathbf{B}_{ij} &= \tilde{\mathbf{X}}_i^\top \hat{\mathbf{b}}_j^p \mathbf{X}_{ij}^\top = \tilde{\mathbf{X}}_i^\top \hat{\mathbf{b}}_j^p (\mathbf{X}_i - \mathbf{X}_j)^\top, \\ W_{ij} &= K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) = K_h((\mathbf{X}_i - \mathbf{X}_j)^\top \tilde{\boldsymbol{\beta}}). \end{aligned}$$

We next formulate the double sum of weighted squares in (4.23) into a traditional weighted least squares. This can be achieved by constructing an $n^2 \times 1$ vector $\mathbf{C}$, an $n^2 \times d_n$ matrix $\mathbb{B}$ and an $n^2 \times n^2$ diagonal matrix $\mathbf{W}$ as follows:

$$\mathbf{C} = (c_{11}, \ldots, c_{n1}, c_{12}, \ldots, c_{n2}, \ldots, c_{1n}, \ldots, c_{nn})^\top, \tag{4.24}$$

$$\mathbb{B} = (\mathbf{B}_{11}, \ldots, \mathbf{B}_{n1}, \mathbf{B}_{12}, \ldots, \mathbf{B}_{n2}, \ldots, \mathbf{B}_{1n}, \ldots, \mathbf{B}_{nn})^\top, \tag{4.25}$$

$$\mathbf{W} = \mathrm{diag}\{W_{11}, \ldots, W_{n1}, W_{12}, \ldots, W_{n2}, \ldots, W_{1n}, \ldots, W_{nn}\}. \tag{4.26}$$

Then, it leads to a minimization problem of a traditional weighted least squares

$$\hat{\boldsymbol{\beta}}^o = \arg\min_{\boldsymbol{\beta}} (\mathbf{C} - \mathbb{B}\boldsymbol{\beta})^\top \mathbf{W} (\mathbf{C} - \mathbb{B}\boldsymbol{\beta}).$$

Following from least squares theory, we compute the preliminary estimator $\hat{\boldsymbol{\beta}}^o = (\hat{\beta}_1^o, \ldots, \hat{\beta}_{d_n}^o)$ by

$$\hat{\boldsymbol{\beta}}^o = (\mathbb{B}^\top \mathbf{W} \mathbb{B})^{-1} (\mathbb{B}^\top \mathbf{W} \mathbf{C}). \tag{4.27}$$

4. Step 4: Using the estimates $\hat{\mathbf{F}}^p = (\hat{\mathbf{f}}_1^p, \ldots, \hat{\mathbf{f}}_n^p)^T$ with $\hat{\mathbf{f}}_j \equiv \{(\hat{\mathbf{a}}_j^p)^\top, (\hat{\mathbf{b}}_j^p)^\top\}^\top$, $j = 1, \ldots, n$, we consider the minimization problem of the locally weighted SCAD function defined as

$$\begin{aligned} Q(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \hat{\mathbf{F}}^p) \;=\; & \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{a}}_j^p - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{b}}_j^p \mathbf{X}_{ij}^\top \boldsymbol{\beta}\}^2 K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) \\ & + \sum_{k=1}^{d_n} p_{\tilde{\lambda}_k}(|\beta_k|), \end{aligned} \tag{4.28}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{d_n})^\top \in \mathbb{R}^{d_n}$, where $\tilde{\boldsymbol{\lambda}} = (\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{d_n})^\top \in \mathbb{R}^{d_n}$ is the vector of tuning (regularization) parameters. According to the idea of local quadratic approximation, with given preliminary estimate $\hat{\boldsymbol{\beta}}^o = (\hat{\beta}_1^o, \ldots, \hat{\beta}_{d_n}^o)$, the penalty functions in (4.28) can be locally approximated by

$$p_{\tilde{\lambda}_k}(|\beta_k|) \approx p_{\tilde{\lambda}_k}(|\hat{\beta}_k^o|) + \frac{1}{2} \frac{p'_{\tilde{\lambda}_k}(|\hat{\beta}_k^o|)}{|\hat{\beta}_k^o|} [(\beta_k^2 - (\hat{\beta}_k^o)^2], \quad \text{for } \beta_k \approx \hat{\beta}_k^o, \tag{4.29}$$

in other words,

$$[p_{\tilde{\lambda}_k}(|\beta_k|)]' \approx \frac{p'_{\tilde{\lambda}_k}(|\hat{\beta}_k^o|)}{|\hat{\beta}_k^o|} \beta_k.$$

Consequently, by (4.28) and (4.29), we define a new objective function

$$\begin{aligned} Q(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \hat{\mathbf{F}}^p, \hat{\boldsymbol{\beta}}^o) &= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{a}}_j^p - \tilde{\mathbf{X}}_i^\top \hat{\mathbf{b}}_j^p \mathbf{X}_{ij}^\top \boldsymbol{\beta}\}^2 \\ &\quad \times K_h(\mathbf{X}_{ij}^\top \hat{\boldsymbol{\beta}}^o) + \frac{1}{2} \sum_{k=1}^{d_n} \frac{p'_{\tilde{\lambda}_k}(|\hat{\beta}_k^o|)}{|\hat{\beta}_k^o|} {\beta_k}^2 \\ &\quad + \bar{C}, \end{aligned} \tag{4.30}$$

where $\bar{C}$ stands for the constant terms when the preliminary estimator $\hat{\boldsymbol{\beta}}^o = (\hat{\beta}_1^o, \ldots, \hat{\beta}_{d_n}^o)$ are provided. Now, we would like to calculate the estimator of $\boldsymbol{\beta}$,denoted by $\hat{\boldsymbol{\beta}}^p = (\hat{\beta}_1^p, \ldots, \hat{\beta}_{d_n}^p)$ such that

$$\hat{\boldsymbol{\beta}}^p = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \hat{\mathbf{F}}^p, \hat{\boldsymbol{\beta}}^o). \tag{4.31}$$

Similar to the way in Step 3, to solve (4.31), we rewrite (4.30) as

$$\begin{aligned} Q(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \hat{\mathbf{F}}^p, \hat{\boldsymbol{\beta}}^o) &= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{i=1}^{n} \{C_{ij} - \mathbf{B}_{ij} \boldsymbol{\beta}\}^2 W_{ij} \\ &\quad + \frac{1}{2} \sum_{k=1}^{d_n} \frac{p'_{\tilde{\lambda}_k}(|\hat{\beta}_k^o|)}{|\hat{\beta}_k^o|} {\beta_k}^2 + \bar{C}, \end{aligned} \tag{4.32}$$

where

$$\begin{aligned} C_{ij} &= Y_i - \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{a}}_j^p, \\ \mathbf{B}_{ij} &= \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{b}}_j^p \mathbf{X}_{ij}^\top = \tilde{\mathbf{X}}_i^\top \hat{\boldsymbol{b}}_j^p (\mathbf{X}_i - \mathbf{X}_j)^\top, \\ W_{ij} &= K_h(\mathbf{X}_{ij}^\top \tilde{\boldsymbol{\beta}}) = K_h((\mathbf{X}_i - \mathbf{X}_j)^\top \tilde{\boldsymbol{\beta}}). \end{aligned}$$

With exactly the same notations (4.24), (4.25) and (4.26) in Step 3, function (4.32) can be written in the matrix notation as

$$Q(\boldsymbol{\beta} \mid \boldsymbol{\lambda}, \hat{\mathbf{F}}^p, \hat{\boldsymbol{\beta}}^o) = (\mathbf{C} - \mathbb{B}\boldsymbol{\beta})^\top \mathbf{W}(\mathbf{C} - \mathbb{B}\boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^\top \Sigma_{\tilde{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\beta}}^o)\boldsymbol{\beta} + \bar{C}, \tag{4.33}$$

where

$$\Sigma_{\tilde{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\beta}}^o) = \operatorname{diag}\left\{\frac{p'_{\tilde{\lambda}_1}(|\hat{\beta}_1^o|)}{|\hat{\beta}_1^o|}, \ldots, \frac{p'_{\tilde{\lambda}_{d_n}}(|\hat{\beta}_{d_n}^o|)}{|\hat{\beta}_{d_n}^o|}\right\}$$

is an $d_n \times d_n$ diagonal matrix. The solution of the minimization problem of (4.33) can be found by

$$\hat{\boldsymbol{\beta}}^p = (\mathbb{B}^\top \mathbf{W}\mathbb{B} + \frac{n}{2}\Sigma_{\tilde{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\beta}}^o))^{-1}(\mathbb{B}^\top \mathbf{W}\mathbf{C}). \tag{4.34}$$

In terms of the identifiability condition assumed in (4.14), the estimator $\hat{\boldsymbol{\beta}}^p$ should be rescaled to satisfy the constraints $||\hat{\boldsymbol{\beta}}^p|| = 1$, $\hat{\beta}_1^p > 0$.

Go back to Step 1 and update the estimate $\tilde{\boldsymbol{\beta}}$ with the scaled $\hat{\boldsymbol{\beta}}^p$ and repeat above four steps until convergence. Concretely, in our implementation, the convergence condition are defined as follows: when the estimates of index parameters from the t-th iteration, $\hat{\boldsymbol{\beta}}^{p(t)}$, satisfy

$$\max\left\{\left|\frac{\hat{\beta}_1^{p(t)} - \hat{\beta}_1^{p(t-1)}}{\hat{\beta}_1^{p(t-1)}}\right|, \ldots, \left|\frac{\hat{\beta}_{d_n}^{p(t)} - \hat{\beta}_{d_n}^{p(t-1)}}{\hat{\beta}_{d_n}^{p(t-1)}}\right|\right\} < \vartheta,$$

where $t = 1, \ldots, T$ is the $t-$th iteration in the procedures, and ϑ, a positive number close to 0, is a threshold, the iterative procedure is convergence. Practically, we set the threshold of convergence be $\vartheta = 0.01$.

When the convergence condition has been fulfilled, we obtain the resulting estimators form the iterative procedure, which are denoted by $\check{\boldsymbol{\beta}} = (\check{\beta}_1, \ldots, \check{\beta}_{d_n})^\top \in \mathbb{R}^{d_n}$ and

$$\check{\mathbf{F}} = \begin{pmatrix} \check{\mathbf{a}}_1, & \check{\mathbf{b}}_1 \\ \vdots & \vdots \\ \check{\mathbf{a}}_n, & \check{\mathbf{b}}_n \end{pmatrix} = (\check{\mathbf{a}}_{(0)}, \ldots, \check{\mathbf{a}}_{(d_n-1)}, \check{\mathbf{b}}_{(0)}, \ldots, \check{\mathbf{b}}_{(d_n-1)}) \in \mathbb{R}^{n \times 2d},$$

respectively.

By applying the foregoing iterative procedure, the resulting estimators of insignificant coefficients and irrelevant index parameters are expected to be shrunk to very small values. Sequentially, in our implementation, we employ thresholds to automatically eliminate the resulting estimators which are smaller than the corresponding thresholds, and thus we obtain the penalised estimator with sparsity.

Specifically, the implementation can be broken down as follows:

1. We assume that the thresholds $\{\theta_1, \theta_2, \theta_3\}$ is small enough and there exists three subsets $\bar{S}_1$, $\bar{S}_2$ and $\bar{S}_3$ of $S_0 \equiv \{0, \ldots, d_n - 1\}$. If $||\check{\mathbf{a}}_{(k)}||$, $k \in S_0$, with respect to the coefficient $f_k(\cdot)$, are smaller than θ_1, we set $||\check{\mathbf{a}}_{(k)}|| = 0$ and the

subset $\bar{S}_3$ as

$$\bar{S}_3 = \{k : k \in S_0, ||\hat{\mathbf{a}}_{(k)}|| > 0\}.$$

2. If $||\check{\mathbf{b}}_{(k)}||$, $k \in \bar{S}_3$, with respect to the first derivative of $f_k(\cdot)$, is smaller than the thresold θ_2, it will be automatically set to 0. Accordingly, we define the subsets $\bar{S}_1$ and $\bar{S}_2$ by

$$\bar{S}_1 = \{k : k \in S_0, ||\hat{\mathbf{a}}_{(k)}|| > 0, ||\check{\mathbf{b}}_{(k)}|| > 0\};$$
$$\bar{S}_2 = \{k : k \in S_0, ||\hat{\mathbf{a}}_{(k)}|| > 0, ||\check{\mathbf{b}}_{(k)}|| = 0\}.$$

   To avoid the abuse of notation, we denote the ultimate sparse estimates for coefficients $f_k(\cdot)$, $k \in S_0$ as

$$\hat{\mathbf{F}} = (\hat{\mathbf{a}}_{(0)}, \ldots, \hat{\mathbf{a}}_{(d_n-1)}, \hat{\mathbf{b}}_{(0)}, \ldots, \hat{\mathbf{b}}_{(d_n-1)}) \in \mathbb{R}^{n \times 2d},$$

   among which, for $k \in \bar{S}_1$, $\hat{\mathbf{a}}_{(k)} = \check{\mathbf{a}}_{(k)}$ and $\hat{\mathbf{b}}_{(k)} = \check{\mathbf{b}}_{(k)}$; for $k \in \bar{S}_2$, $\hat{\mathbf{a}}_{(k)} = \check{\mathbf{a}}_{(k)}$ and $\hat{\mathbf{b}}_{(k)} = 0$; the rest $\hat{\mathbf{a}}_{(k)}$ and $\hat{\mathbf{b}}_{(k)}$ equal to $\mathbf{0}$.

3. Similarly, assuming that $\tilde{S}$ is the subset of $S \equiv \{1, \ldots, d_n\}$, if $\left|\check{\beta}_k\right| < \theta_3$, $k \in S$, we set $\check{\beta}_k = 0$ and define the subset $\tilde{S}$ as

$$\tilde{S} = \{k : k \in S, \check{\beta}_k = 0\}.$$

   Then, $\check{\boldsymbol{\beta}}$ becomes a sparse estimates for index parameter $\boldsymbol{\beta}$. By rescaling $\check{\boldsymbol{\beta}}$ to satisfy the constraints $||\boldsymbol{\beta}|| = 1$ and

$\beta_1 > 0$, we obtain the ultimate sparse estimates for index parameter $\boldsymbol{\beta}$, which are denoted by

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{d_n})^\top.$$

It is necessary to point out that we have suppressed the dependency of the ultimate estimates on $\tilde{\boldsymbol{\lambda}}$. Furthermore, it can be seen that in the foregoing computational algorithms, we shrink the irrelevant components of the underlying model to zero only after the iterative procedure completed. This implementation leads to a "double check" mechanism which works as follows: if after an iteration a coefficient or an index parameter is shrunken to be insignificant, it still has an opportunity to be reselected into the model in the following iteration. Thanks to this mechanism, our algorithm can overcome the main drawback in typical local quadratic approximation, which is that once a coefficient is lessened to zero, it will remain at zero. Meanwhile, since we do not eliminate the insignificant components in each iteration, the algorithm is not very sensitive to the choice of initial values, namely, the choice of initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. We will show the ultimate estimates are not sensitive to the choice of initial estimate in our simulation study in Section 7.1.

4.2.3 Modification of the proposed algorithm

In the proposed iterative procedure, we notice that in solutions (4.27) and (4.34) within Step 3 and Step 4, respectively, the memory required for execution grows at order $O(n^2 d_n)$. It is acceptable to conduct the proposed algorithm in the modest dimensional models, but in the high dimensionality, care shall be taken from a computational point of view to avoid exceeding the limitations of memory.

To address this issue, in practical implementation, we consider a slight modification on computing two quantities $\mathbb{B}^\top \mathbf{W} \mathbb{B}$ and $\mathbb{B}^\top \mathbf{W} \mathbb{C}$ involving in both (4.27) and (4.34). Based on the notation (4.24) - (4.26), we first denote an $n \times d_n$ matrix $\mathscr{B}_j$, an $n \times 1$ vector $\mathscr{C}_j$ and an $n \times n$ matrix $\mathscr{W}_j$ as follows:

$$\mathscr{B}_j = \begin{pmatrix} B_{1j} \\ \vdots \\ B_{nj} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{X}}_1^\top \hat{\boldsymbol{b}}_j^p (\mathbf{X}_1 - \mathbf{X}_j)^\top \\ \vdots \\ \tilde{\mathbf{X}}_n^\top \hat{\boldsymbol{b}}_j^p (\mathbf{X}_n - \mathbf{X}_j)^\top \end{pmatrix},$$

$$\mathscr{C}_j = \begin{pmatrix} C_{ij} \\ \vdots \\ C_{ij} \end{pmatrix} = \begin{pmatrix} Y_1 - \tilde{\mathbf{X}}_1^\top \hat{\boldsymbol{a}}_j^p \\ \vdots \\ Y_n - \tilde{\mathbf{X}}_n^\top \hat{\boldsymbol{a}}_j^p \end{pmatrix},$$

and

$$\mathscr{W}_j = \text{diag}\{W_{1j}, \ldots, W_{nj}\} = \text{diag}\{K_h(\mathbf{X}_{1j}^\top \tilde{\boldsymbol{\beta}}), \ldots, K_h(\mathbf{X}_{nj}^\top \tilde{\boldsymbol{\beta}})\}.$$

Then, (4.27) and (4.34) can be calculated using the following equations

$$\mathbb{B}^\top \mathbf{W} \mathbb{B} = \sum_{j=1}^{n} \mathscr{B}_j^\top \mathscr{W}_j \mathscr{B}_j, \tag{4.35}$$

$$\mathbb{B}^\top \mathbf{W} \mathbb{C} = \sum_{j=1}^{n} \mathscr{B}_j^\top \mathscr{W}_j \mathscr{C}_j. \tag{4.36}$$

According to (4.35) and (4.36), we actually figure out (4.27) by

$$\hat{\boldsymbol{\beta}}^o = (\sum_{j=1}^{n} \mathscr{B}_j^\top \mathscr{W}_j \mathscr{B}_j)^{-1} (\sum_{j=1}^{n} \mathscr{B}_j^\top \mathscr{W}_j \mathscr{C}_j). \tag{4.37}$$

Similarly, in practice, we compute (4.34) by

$$\hat{\boldsymbol{\beta}}^p = (\sum_{j=1}^{n} \mathscr{B}_j^\top \mathscr{W}_j \mathscr{B}_j + \frac{n}{2} \Sigma_{\tilde{\boldsymbol{\lambda}}}(\hat{\boldsymbol{\beta}}^o))^{-1} (\sum_{j=1}^{n} \mathscr{B}_j^\top \mathscr{W}_j \mathscr{C}_j). \tag{4.38}$$

In (4.37) and (4.38) the space complexity are reduced to $O(nd_n)$, which is acceptable even in the high dimensional situation. Moreover, as the trade-off between the space complexity and time cost should be taken into consideration, we do not apply an algorithm to reduce the space complexity to $O(1)$. The reason is, in programming, the time cost of most algorithms applied for matrix multiplication is less than of the nested loops.

CHAPTER 5

SELECTION OF BANDWIDTH AND TUNING PARAMETER

In this chapter, we will explore the selection of bandwidth and tuning parameter. Both of them play a very important role in our proposed approach.

Specifically, the objective function (4.13) in our algorithm actually contains two parts which can be intuitively expressed as follows:

$$Q(\mathbf{F}, \boldsymbol{\beta}) = \mathscr{L}(\mathbf{F}, \boldsymbol{\beta}) + P(\mathbf{F}, \boldsymbol{\beta}),$$

where $\mathscr{L}(\cdot)$ is the loss function, which measures the fitting of the model and $P(\cdot)$ refers to penalty functions or regularization terms, which control the complexity of the model. In our proposed method, the choice of bandwidth is rather crucial to the model fitting and the SCAD penalty applied to

control the model complexity relies on the proper choice of tuning parameters. Moreover, both bandwidth and tuning parameter will simultaneously impact the model selection and estimation, since both of these two hyper-parameters can control the trade-off between the bias and variance in resulting estimators.

We will explore the selection of bandwidth in Section 5.1 and address how to choose tuning parameter in Section 5.2.

5.1 Bandwidth selection

In this section, we discuss the choice of the bandwidth h involved in the estimation of $f_k(\cdot)$ and $\boldsymbol{\beta}$ in SIVC model.

In order to select the bandwidth in a particular scale, instead of directly tuning the global bandwidth h, we will tune the percentage of the whole range of the estimated index covered by the global bandwidth, which is defined as follows

$$H = \frac{h}{\max\{\hat{Z}_1,\ldots,\hat{Z}_n\} - \min\{\hat{Z}_1 - \hat{Z}_n\}} \times 100\%$$

where h is the value of global bandwidth and $\hat{Z}_i = \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}$, $i = 1,\ldots,n$, for a given estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

5.1.1 Sensitivity to the choice of bandwidth

By defining the relative bandwidth H, we are going to employ a data-driven method to evaluate the performance of the estimation with a sequence of bandwidth parameters from 0 to 100%.

Let $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$ denote the observations, it states in Fan and Gijbels(1996) that a theoretical optimal bandwidth is obtained by minimizing the conditional Mean Square Error (MSE) given $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ or the conditional weighted Mean Integrated Square Error (MISE) given $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$. Accordingly, the criteria used for assessing the performance of the resulting estimates are their MSE and Relative MISE.

Specifically, we employ MSE to measure the goodness of the estimated index parameter $\hat{\boldsymbol{\beta}}$, which is defined as follows:

$$\text{MSE} = \frac{1}{dL} \sum_{l=1}^{L} \sum_{k=1}^{d_n} (\hat{\beta}_k^l - \beta_k)^2,$$

where $\hat{\beta}_k^l$ is either the unpenalised estimator or the penalised estimator from the l-th replication in a simulation, β_k is the true index parameter; and we evaluate the goodness of estimators with respect to coefficients $f_k(\cdot)$, $k = 0, \ldots, d_n - 1$ in terms of the relative MISE (RMISE), which can be approximated by

$$\text{RMISE} \approx \frac{1}{L} \sum_{l=1}^{L} [\frac{\sum_{k=0}^{d_n-1} \sum_{j=1}^{n} (\hat{f}_k^l(z_j) - f_k(z_j))^2}{\sum_{k=0}^{d_n-1} \sum_{j=1}^{n} f_k(z_j)^2}],$$

where $\hat{f}_k^l(\cdot), k = 0,1,\ldots,d_n - 1, l = 1,2,\ldots,L$, is either the unpenalised estimator or the penalised estimator of the $k-$th functional coefficient in the $l-$th replication and $z_j = x_j^T \hat{\boldsymbol{\beta}}, j = 1,\ldots,n$.

Now, we use the grid-search approach based on a simulation study to illustrate the relationship between different bandwidth and the estimation accuracy, namely, to explore the sensitivity of the estimation accuracy to the choice of bandwidth. Consider the following example

$$Y_i = 2\cos(0.5\pi Z_i) + X_{i1} + 4\exp(-Z_i^2)X_{i2} + \varepsilon_i, \qquad (5.1)$$
$$\text{with } Z_i = \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\beta} = \frac{1}{3}(2X_{i1} + 2X_{i2} + X_{id}),$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^{\mathrm{T}}$, for $i = 1,\ldots,n$, are normally distributed independent random vectors and noise ε_i are independent N$(0,1)$ random variables. The regression models are based on the form (4.2) with $\boldsymbol{\beta} = \frac{1}{3}(2,2,0,\ldots,1)^{\mathrm{T}}$.

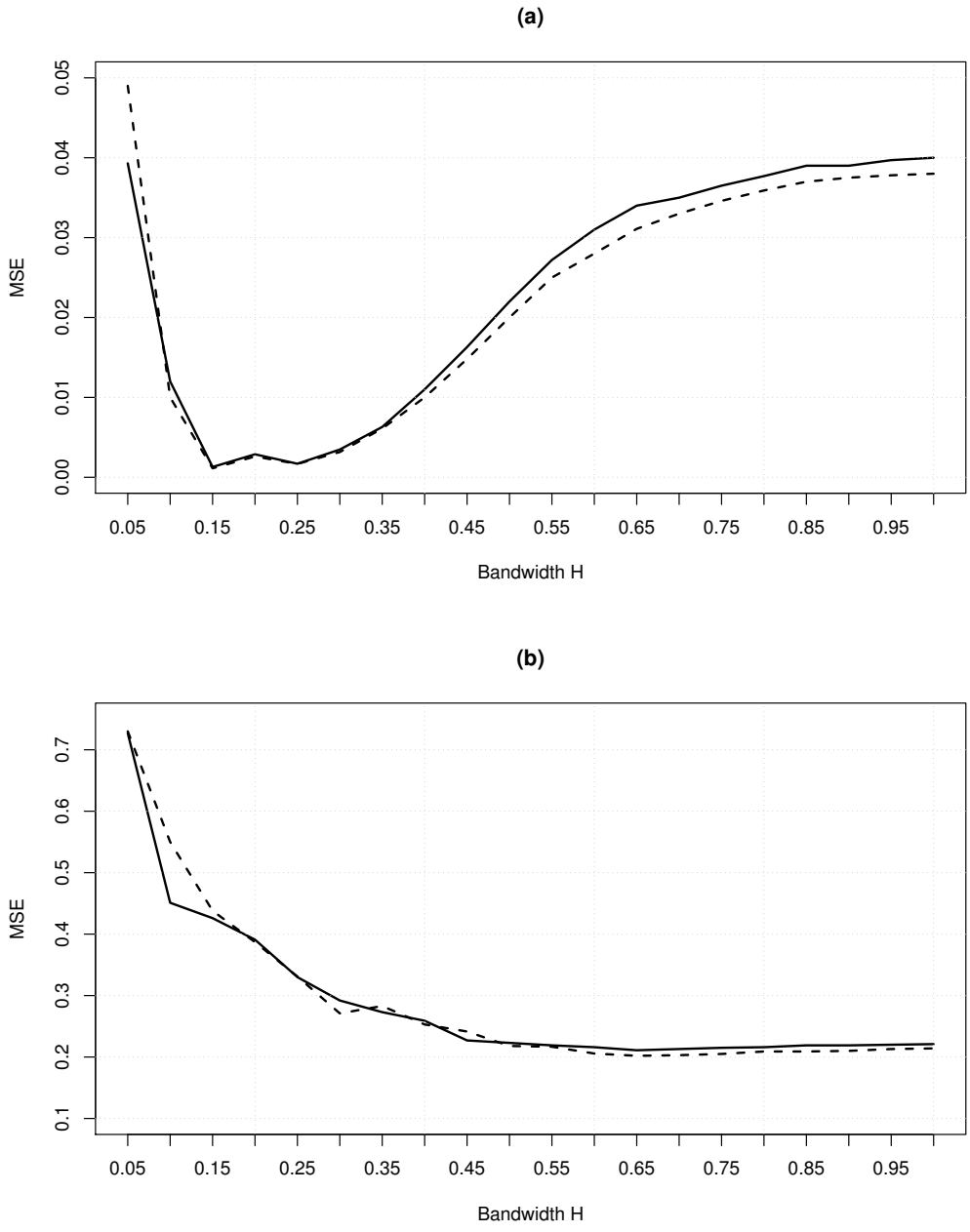
We first consider the underlying regression model with a modest dimension that $d = 7$. Then, we conduct simulations on the model by applying the unpenalised approach developed in Section 3.2.2 and the proposed penalised approach in Section 4.2.2, respectively. We use a uniform tuning parameter $\lambda_f = 4$ for selecting the coefficients $f(\cdot)_k$ and another uniform tuning parameter $\lambda_\beta = 40$ for obtaining the penalised estimates of index parameters. The kernel function $k(\cdot)$ we used through out this section and the following numerical analyses

is Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$. We simulate in 1000 datasets, each with the sample size $n = 600$.

The simulation results about how the choice of bandwidth impact the estimation of index parameter $\boldsymbol{\beta}$ and coefficient $f_k(\cdot)$ are given in Figure 5.1 and Figure 5.2, respectively.

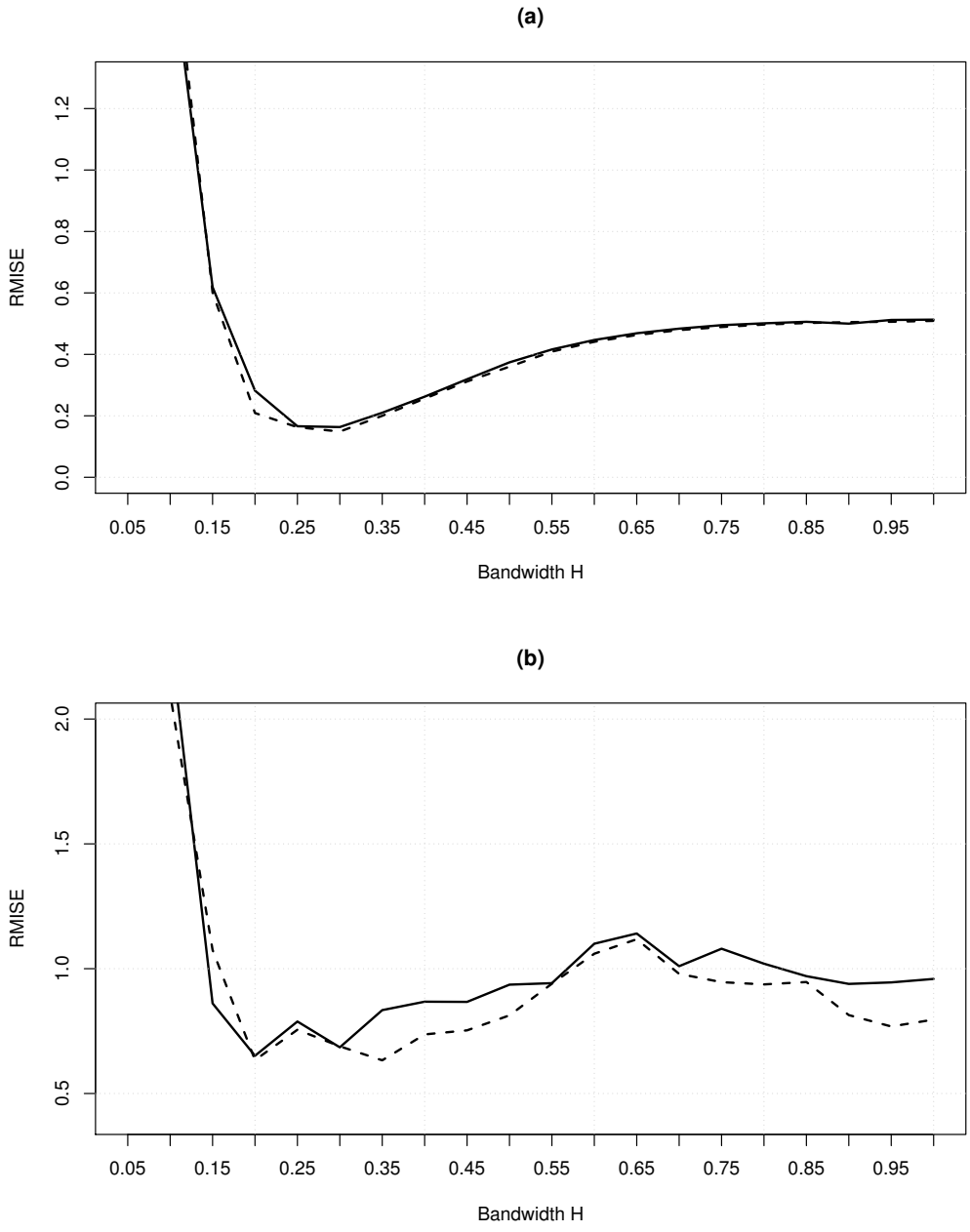
The finding from the results is threefold. Firstly, the choice of bandwidth is essential, since in all the cases, either MSE or RMISE can be remarkably reduced to a reasonable value by a careful choice of bandwidth. Secondly, there indeed exists the optimal bandwidth for penalised estimate of $\boldsymbol{\beta}$, which is inside the range $(0.15, 0.35)$ and the optimal bandwidth for penalised estimates of $f_k(\cdot)$ locates in the range $(0.25, 0.35)$. Thirdly, the penalised estimators perform significantly better than the unpenalised estimators and are more sensitive to the choice of bandwidth.

Figure 5.1: Sensitivity of MSE to bandwidth H



NOTE: *Simulation results: (a) sensitivity of MSE of penalised estimates to H; (b) sensitivity of MSE of unpenalised estimates to H. In both cases: solid line, estimate on the underlying model with noise ε_i; dashed line, estimate on the underlying model without noise.*

Figure 5.2: Sensitivity of RMISE to bandwidth H



NOTE: *Simulation results: (a) sensitivity of RMISE of penalised estimates to H; (b) sensitivity of RMISE of estimates of functional coefficients to H; (c) sensitivity of MSE of the estimates of index parameter to H. In both cases: solid line, estimate on underlying model with noise ε_i; dashed line, estimate on underlying model without noise.*

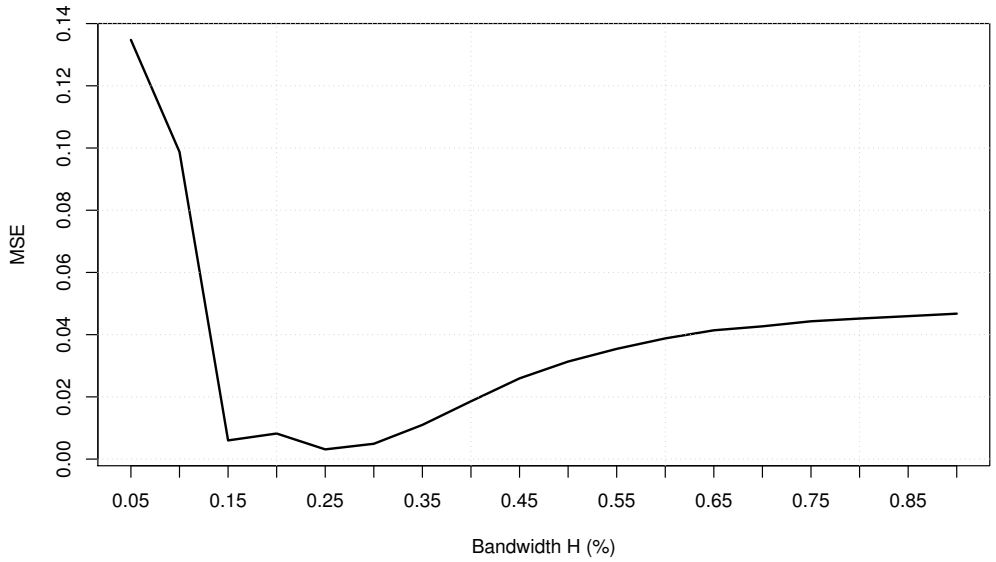
Furthermore, we also execute a simulation on the underlying model with dimension $d = 20$ by merely using penalised approach to demonstrate the sensitivity of estimation accuracy to the bandwidth in the high-dimensional situation. Here, we set an uniform tuning parameter for selecting the coefficients $f(\cdot)_k$ as $\lambda_f = 5$ and the other uniform tuning parameter for selecting index parameters be $\lambda_\beta = 50$. We also conduct simulation with sample size $n = 600$ in a total of 1000 replications. The results are reported in Figure 5.3 and 5.3. From these two figures, we remark that in the high-dimensional model, both the optimal choice of bandwidth and the sensitivity to the choice of this hyper-parameter are fairly similar to the situation in the modest-dimensional model. The optimal bandwidth for penalised estimates of $\boldsymbol{\beta}$ exists in range $(0.15, 0.35)$ and the optimal bandwidth for penalised estimates of $f_k(\cdot)$ is in the range $(0.25, 0.35)$.

5.1.2 Bandwidth selection in practical implementation

Since in the real dataset, the true parameters are unknown, the MSE-criterion or RMISE-criterion discussed in Section 5.1 is unable to be applied.

The cross-validation is a possible way to select the bandwidth. Wu *et al.*(1998) proposed to use this statistic to choose

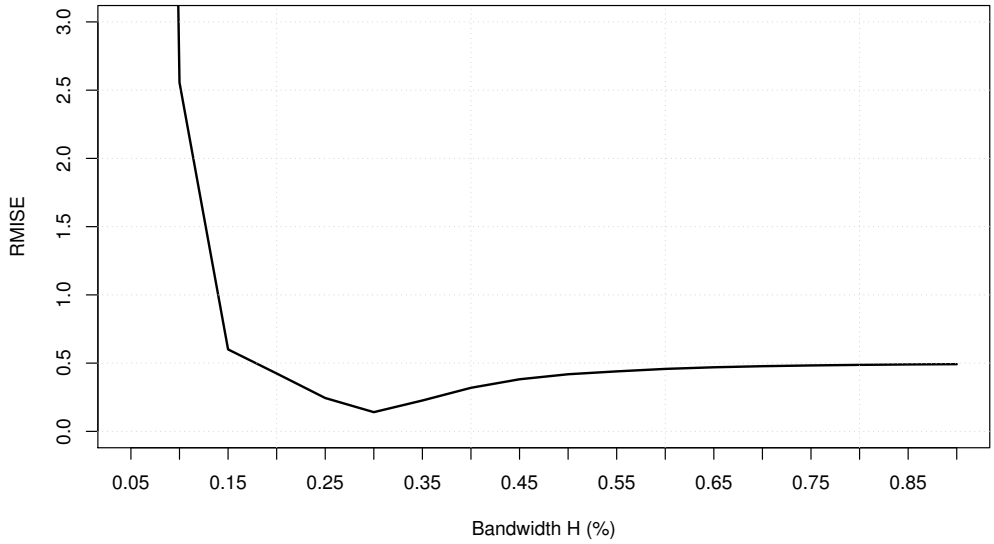
Figure 5.3: Sensitivity of MSE to bandwidth H in high-dimensional situation



the bandwidth.

However, it has been shown in Yang (2005) that cross-validation is asymptotically equivalent to the Akaike information criterion (AIC) and hence they share similar performance. Meanwhile, it is also known that the hyper-parameters selected by AIC may lead to overfitting (Shao, 1997). Therefore, the cross validation cannot consistently identify the optimal bandwidth, whose choice always leads to a relatively high variance in the resulting estimates. The other shortcoming is that the computational expense of a grid-search approach based on cross validation is very high. From our experience, in the high dimensional situation, the parallel computing should

Figure 5.4: Sensitivity of RMISE to bandwidth H in high-dimensional situation



be applied to speed up the computation of cross-validation.

As it is supported by the empirical evidence in Section 5.1.1 that the ultimate estimates from our proposed approach are not very sensitive to the choice of the bandwidth as long as H is chosen to be within a reasonable range, we recommend to follow the idea of Li, Ke and Zhang (2015) to choose the bandwidth as $H = 0.6(d_n/n)^{0.2}$ in the practical implementation.

5.2 Selection of tuning parameter

Selection of the tuning (regularization) parameters is essential to the procedure proposed in Section 4 for the purpose of model selection and structure specification. The tuning parameter vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \cdots, \lambda_{2d_n-1}, \tilde{\lambda}_1, \cdots, \tilde{\lambda}_{d_n})$ is of dimension $3d_n$, and to simultaneously choose a total of $3d_n$ tuning parameters is very challenging. Therefore, we consider a 2-dimensional problem about $\boldsymbol{\lambda} = (\lambda_f, \lambda_\beta) \in \mathbb{R}^2$, which can be selected by the generalized information criterion (GIC) proposed by Fan and Tang (2013).

Since the non-zero coefficients of the models may consist of both varying coefficients and the constant coefficients, we need to work out how many constant parameters each functional parameter amounts to. We follow the idea of Cheng, Zhang and Chen (2009), which suggests that when sample size n is sufficiently large, in the local linear fitting, an unknown functional parameter approximately amounts to $1.028571h^{-1}$ constant parameters when Epanechnikov kernel is applied. Consequently, the tuning parameters $\boldsymbol{\lambda} = (\lambda_f, \lambda_\beta) \in \mathbb{R}^2$ can be selected according to the following criterion

$$\mathrm{GIC}(\lambda_f, \lambda_\beta) = \log(\mathrm{RSS}_{\boldsymbol{\lambda}}) + \frac{a_n}{n} \times (\hat{df} + 1.028571h^{-1}\tilde{df}), \quad (5.2)$$

where

$$a_n = \log(\log(n))\log(1.028571h^{-1}d_n + d_n),$$

$\hat{df}$ is the number of significant constant parameters, $\tilde{df}$ is the number of the significant functional parameters and RSS_λ is defined as

$$\text{RSS}_{\boldsymbol{\lambda}} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_{k=0}^{d_n}\hat{f}_{k(\lambda_f)}(\mathbf{x}_i^T\hat{\beta}_{(\lambda_\beta)})x_{ik}\}^2.$$

Then, the tuning parameter $\boldsymbol{\lambda} = (\lambda_f, \lambda_\beta) \in \mathbb{R}^2$ is obtained by

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\lambda_f,\lambda_\beta} \text{GIC}(\lambda_f, \lambda_\beta).$$

In the practical implementation, we apply an iterative algorithm to get the optimal tuning parameters.

1. *Step 1:* by specifying an initial value of λ_β^0 and based on GIC in (5.2), we select the tuning parameter $\lambda_f \in \mathbb{R}^1$ through

   $$\text{GIC}_{\lambda_f} = \log(\text{RSS}_{\lambda^0}) + \frac{\bar{a}_n}{n} \times (\hat{df}_f + 1.028571h^{-1}\tilde{df}_f),$$

   where

   $$\bar{a}_n = \log\{\log(n)\}\log(1.028571h^{-1}d_n),$$

   $\hat{df}_f$ is the number of relevant covariates with constant coefficients, $\tilde{df}_f$ is the number of relevant covariates with varying coefficients and RSS_{λ^0} is

   $$\text{RSS}_{\lambda^0} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_{k=0}^{d_n}\hat{f}_{k(\hat{\lambda}_f)}(\mathbf{x}_i^T\hat{\beta}_{(\lambda_\beta^0)})x_{ik}\}^2.$$

Then, we determine the tuning parameter $\lambda_f \in \mathbb{R}^1$ by

$$\hat{\lambda}_f = \arg\min_{\lambda_f} \mathrm{GIC}_{\lambda_f}.$$

2. *Step 2:* by updating $\hat{\lambda}_f$, the tuning parameter of index parameter $\lambda_\beta \in \mathbb{R}^1$ is selected by

$$\hat{\lambda}_\beta = \arg\min_{\lambda_\beta} \mathrm{GIC}_{\lambda_\beta},$$

with

$$\mathrm{GIC}_{\lambda_\beta} = \log(\mathrm{RSS}_{\boldsymbol{\lambda}^1}) + n^{-1}\log(\log(n))\log(d_n) \times df_{\lambda_\beta},$$

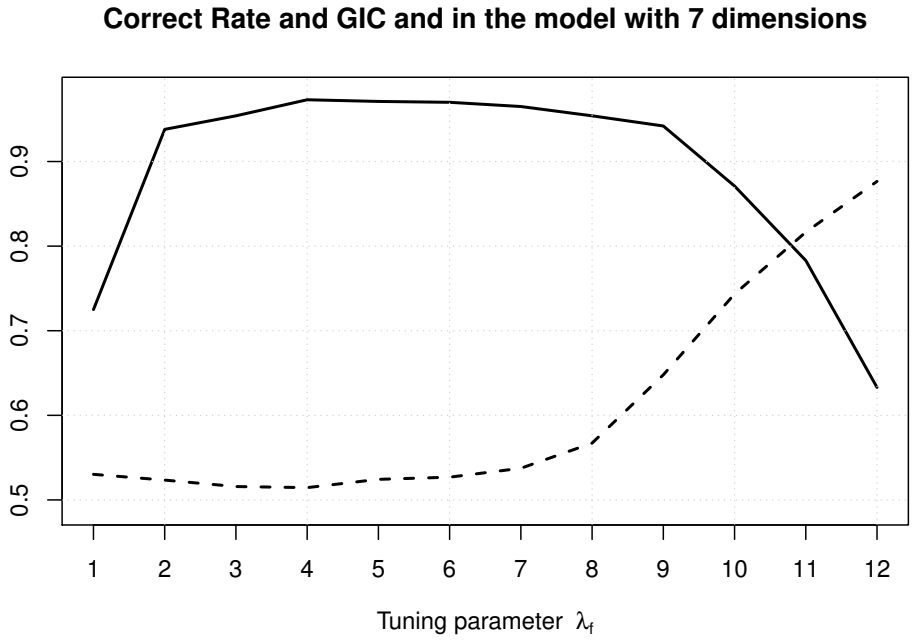
where df_{λ_β} is the number of significant index parameters and RSS_{λ^1} is

$$\mathrm{RSS}_{\lambda^1} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \sum_{k=0}^{d_n} \hat{f}_{k(\hat{\lambda}_f)}(\mathbf{x}_i^T \hat{\beta}_{(\hat{\lambda}_\beta)})x_{ik}\}^2.$$

Now, we use a simulation study to illustrate the accuracy of GIC and show that the sensitivity to the choice of λ_f and λ_β are different. We introduce a criterion termed "Correct Rate" to measure the performance of model selection. Whenever an estimated model is exactly the true model that includes all the relevant elements but does not contain any irrelevant components, we classify it as a "correct model". The ratio of obtaining the "correct model" from all the replications in the simulation is defined as the "Correct Rate".

We first fix $\lambda_\beta = 40$ and search the performance of GIC with respect to λ_f. We conduct simulations on the same simulated example (5.1) described in Section 5.1.1 with dimensions $d = 7$ and $d = 20$. We execute the simulation with sample size $n = 600$ in a total of 1000 replications. The simulated results are reported in Table 5.1 and also depicted in Figure 5.5 and Figure 5.6.

Figure 5.5: The performance of GIC_{λ_f} in the modest dimensionality



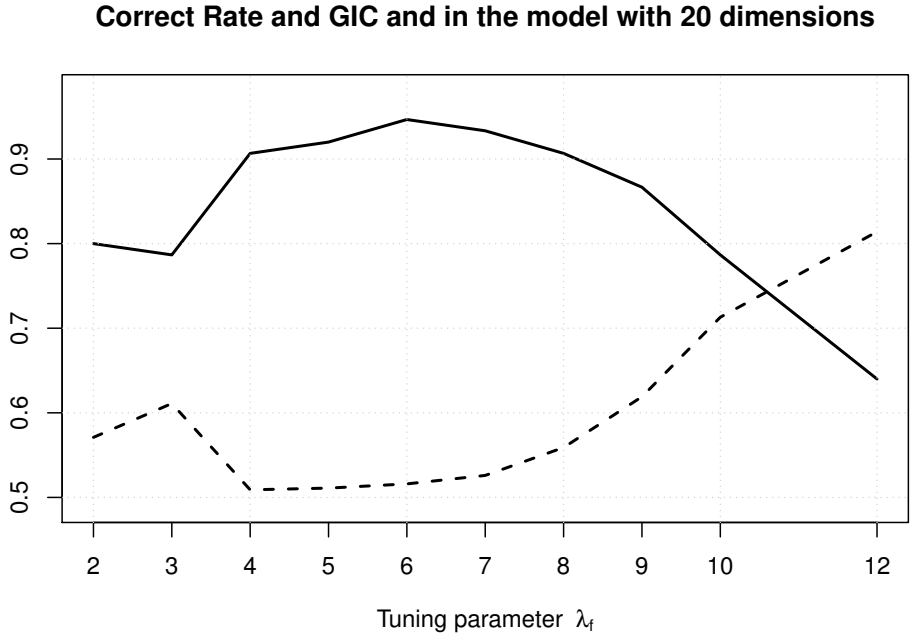
NOTE: *The solid curve indicates the Correct Rate and the dashed curve refers to the GIC with respect to different tuning parameters* λ_f.

Table 5.1: The performance of GIC with respect to λ_f

λ_f	$d = 7$		$d = 20$	
	GIC_{λ_f}	CR	GIC_{λ_f}	CR
1	0.5302	0.725	0.8153	0.223
2	0.5234	0.938	0.5727	0.802
3	0.5158	0.954	0.6277	0.788
4	0.5145	0.973	0.5785	0.905
5	0.5244	0.971	0.5702	0.921
6	0.5269	0.970	0.5451	0.944
7	0.5375	0.965	0.5635	0.934
8	0.5672	0.954	0.6226	0.907
9	0.6479	0.942	0.6534	0.862
10	0.7434	0.876	0.7370	0.783
12	0.8765	0.633	0.8682	0.641
15	1.0171	0.400	1.0188	0.457
20	1.2202	0.358	1.2106	0.388
30	1.4530	0.281	1.4244	0.224

NOTE: *The label "CR" represents "Correct Rate"; d=7 and d=20 refer to the simulation conducted on the model with 7 dimensions and 20 dimensions, respectively.*

Figure 5.6: GIC with respect to different tuning parameters λ_f in the high dimensionality



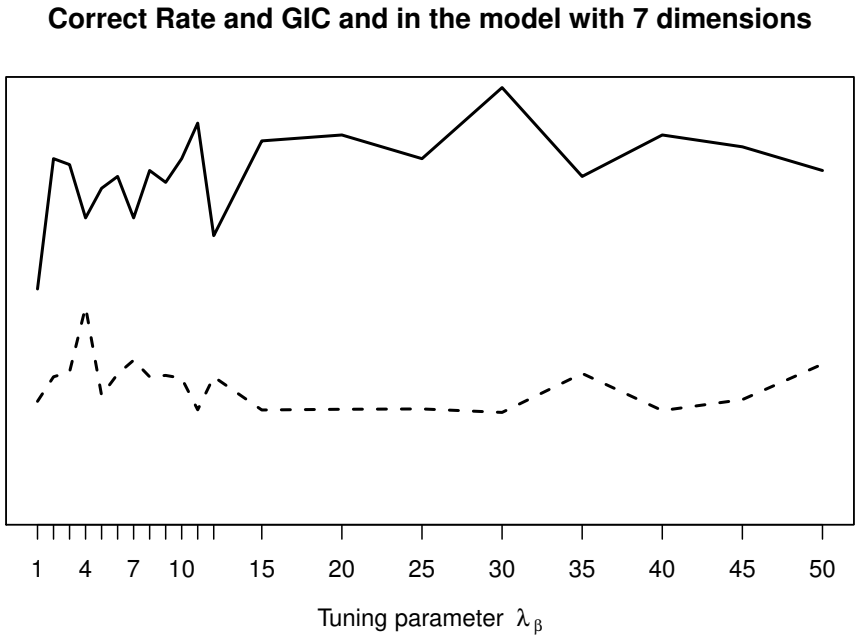
NOTE: *The solid curve indicates the Correct Rate and the dashed curve refers to the GIC with respect to different tuning parameters* λ_f.

It can be seen from Table 5.1 that GIC is able to precisely detect the optimal tuning parameter λ_f and from Figure 5.5 and Figure 5.6, we notice that this criterion can consistently identify the pattern of the corresponding correct rates to a sequence of different tuning parameters. All the simulated results corroborate that the GIC works quite well in both modest dimensional and the high dimensional models.

Then, by fixing λ_f, we conduct another simulations to obtain the performance of GIC concerning λ_β in the model with

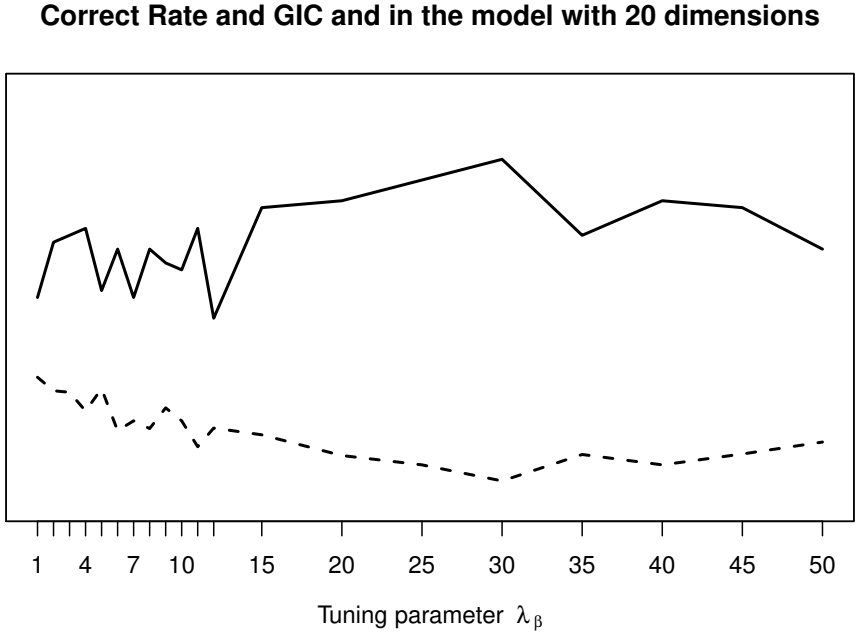
7 dimensions and 20 dimensions, respectively. We intuitively present the simulated results in Figure 5.7 and Figure 5.8 and reported the details in Table 5.2.

Figure 5.7: GIC with respect to different tuning parameters λ_β in the modest dimensionality



NOTE: *The solid curve depicts the Correct Rate and the dashed curve indicates the GIC concerning different tuning parameters* λ_β.

Figure 5.8: GIC with respect to different tuning parameters λ_β in the high dimensionality



NOTE: *The solid curve depicts the Correct Rate and the dashed curve indicates the GIC concerning different tuning parameters* λ_β.

We can see from the Table 5.2 that the goodness of model selection is not very sensitive to the choice of λ_β but the tuning parameter λ_β determined by the $\text{GIC}_{\lambda_\beta}$ is precisely the optimal one. In addition, we notice that that the sensitivity to λ_β is obviously different from the sensitivity to the λ_f, which can be regarded as an empirical evidence that considering the 2-dimensional problem of tuning parameter is more reasonable than only choosing one globally unified tuning parameter.

To conclude, the aforementioned simulation results corrob-

Table 5.2: The performance of $\mathrm{GIC}_{\lambda_\beta}$

λ_β	d=7		d=20	
	$\mathrm{GIC}_{\lambda_\beta}$	CR	$\mathrm{GIC}_{\lambda_\beta}$	CR
1	0.4055	0.947	0.3846	0.931
2	0.4091	0.969	0.3820	0.938
3	0.4097	0.968	0.3817	0.939
4	0.4191	0.959	0.3781	0.940
5	0.4063	0.964	0.3824	0.931
6	0.4094	0.966	0.3743	0.937
7	0.4114	0.959	0.3762	0.931
8	0.4090	0.967	0.3747	0.937
9	0.4092	0.965	0.3787	0.935
10	0.4088	0.969	0.3762	0.934
11	0.4042	0.975	0.3712	0.940
12	0.4088	0.956	0.3748	0.927
15	0.4042	0.972	0.3735	0.943
20	0.4043	0.973	0.3695	0.944
25	0.4044	0.969	0.3677	0.947
30	0.4038	0.981	0.3646	0.949
35	0.4095	0.966	0.3697	0.939
40	0.4041	0.973	0.3676	0.944
45	0.4057	0.971	0.3698	0.943
50	0.4108	0.965	0.3721	0.937

NOTE: *The label "CR" represents "Correct Rate"; d=7 and d=20 refer to the simulation conducted on the model with 7 dimensions and 20 dimensions, respectively.*

orates that the GIC can identify the true model consistently.

CHAPTER 6

ASYMPTOTIC PROPERTIES

In this chapter, we are going to present the asymptotic properties of the model selection and estimation procedure in Section. Before that, we will list technical conditions needed for the asymptotic properties of our proposed method in Section 6.1. The detailed proofs of these theoretical results are given in Chapter 9.

Before presenting the theoretic results, we first rewrite the objective function (4.13), because the identification of constant functions has to be based on the estimates of their derivatives, which are somewhat unreliable. We, therefore, propose the

following alternative to (4.13):

$$\begin{aligned} Q_n(\mathbf{a},\mathbf{F},\mathbf{B},\boldsymbol{\beta}|\boldsymbol{\lambda}) \;=\; & \frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{Y_i-\tilde{\mathbf{X}}_i^\top(\mathbf{a}+\mathbf{f}_j)-(\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_j\}^2 \\ & \times K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) \\ & +\frac{1}{n^2}\sum_{i=1}^{n}\{Y_i-\tilde{\mathbf{X}}_i^\top\mathbf{a}-(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1\}^2 \\ & \times K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})+\sum_{k=0}^{d-1}p_{\lambda_k}(|a_k|+|\mathbf{f}_{(k)}|) \\ & +\sum_{k=0}^{d-1}p_{\lambda_{k+d}}(|\mathbf{f}_{(k)}|)+\sum_{k=1}^{d}p_{\tilde{\lambda}_k}(|\beta_k|), \end{aligned} \tag{6.1}$$

which is minimized with respect to $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_d)^\top\in\mathbb{R}^d$, $\mathbf{a}=(a_0,a_1,\cdots,a_{d-1})^\top$, $\mathbf{B}=(b_{jk})=[\mathbf{b}_1,\cdots,\mathbf{b}_n]^\top=[\mathbf{b}_{(0)},\cdots,\mathbf{b}_{(d-1)}]\in\mathbb{R}^{n\times d}$, and $\mathbf{F}=(f_{jk})=[\mathbf{f}_2,\cdots,\mathbf{f}_n]^\top=[\mathbf{f}_{(0)},\cdots,\mathbf{f}_{(d-1)}]\in\mathbb{R}^{(n-1)\times d}$, where h_n is the smoothing parameter, such that $h_n\to 0$ as $n\to\infty$. Write the minimiser as $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{a}}=(\hat{a}_0,\cdots,\hat{a}_{d-1})^\top$, $\hat{\mathbf{B}}$ and $\hat{\mathbf{F}}$, respectively. We note that the columns of $\hat{\mathbf{F}}$, i.e., $\{\hat{\mathbf{f}}_{(k)},k\in S_0\}$, are respectively estimates of

$$\mathbf{f}_{(k)}=(f_k(\mathbf{X}_1^\top\boldsymbol{\beta}),\cdots,f_k(\mathbf{X}_n^\top\boldsymbol{\beta}))^\top,\quad k\in S_0;$$

while $\{\hat{\mathbf{b}}_{(k)},k\in S_0\}$, the columns of $\hat{\mathbf{B}}$, are respectively estimates of the derivatives of the functional coefficients

$$\dot{\mathbf{f}}_{(k)}^n=h_n(\dot{f}_k(\mathbf{X}_1^\top\boldsymbol{\beta}),\cdots,\dot{f}_k(\mathbf{X}_n^\top\boldsymbol{\beta}))^\top,\quad k\in S_0.$$

The function $Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\lambda})$ in (6.1) is formed in such manner so that, for $k=0,\cdots,d-1$, $\hat{a}_k$ acts as an estimate of $f_k(\mathbf{X}_1^\top\boldsymbol{\beta}_0)$,

while $\hat{\mathbf{f}}_{(k)}$, as an estimate of $(f_k(\mathbf{X}_2^\top \boldsymbol{\beta}_0) - f_k(\mathbf{X}_1^\top \boldsymbol{\beta}_0), \cdots, f_k(\mathbf{X}_n^\top \boldsymbol{\beta}_0) - f_k(\mathbf{X}_1^\top \boldsymbol{\beta}_0))^\top$, which is a vector of zeros if $f_k(.)$ is a constant function. With the penalty imposed on $|\mathbf{f}_{(k)}|$, sparse solutions (estimates) could be induced for these vectors. Because of a slight abuse of notation, we emphasise that notations $\mathbf{a}$, $\mathbf{F}$ and $\mathbf{B}$ are defined differently from the previous chapters and we merely use the newly defined notations in this chapter and the Chapter 9. Then, based on (6.1) we rewrite the proposed iterative procedure with brevity. Start with an initial estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_0$.

- *Step 1:* Minimize the quantity below with respect to $\mathbf{a}$, $\mathbf{F}$ and $\mathbf{B}$:

$$
\begin{aligned}
Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\tilde{\boldsymbol{\beta}},\boldsymbol{\lambda}) \;:=\; & \frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a}+\mathbf{f}_j) - (\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_j\}^2 \\
& K_{h_n}(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}}) \\
& + \frac{1}{n^2}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top\mathbf{a} - (\mathbf{X}_{i1}^\top\tilde{\boldsymbol{\beta}}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1\}^2 \\
& \times K_{h_n}(\mathbf{X}_{i1}^\top\tilde{\boldsymbol{\beta}}) \\
& + \sum_{k=0}^{d-1} p_{\lambda_k}(|\alpha_k| + |\mathbf{f}_{(k)}|) + \sum_{k=0}^{d-1} p_{\lambda_{k+d}}(|\mathbf{f}_{(k)}|);
\end{aligned}
\tag{6.2}
$$

  denote the minimizer as $\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}})$, $\hat{\mathbf{F}}(\tilde{\boldsymbol{\beta}}) = (\hat{\mathbf{f}}_2(\tilde{\boldsymbol{\beta}}), \cdots, \hat{\mathbf{f}}_n(\tilde{\boldsymbol{\beta}}))^\top$, and $\hat{\mathbf{B}}(\tilde{\boldsymbol{\beta}}) = (\hat{\mathbf{b}}_1(\tilde{\boldsymbol{\beta}}), \cdots, \hat{\mathbf{b}}_n(\tilde{\boldsymbol{\beta}}))^\top$, respectively. Note that for ease of notation, we have suppressed the dependency of these quantities on complexity parameter vector $\boldsymbol{\lambda}$.

- *Step 2:* Minimize

$$\frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{Y_i-\tilde{\mathbf{X}}_i^\top(\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}})+\hat{\mathbf{f}}_j(\tilde{\boldsymbol{\beta}}))-(\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_j(\tilde{\boldsymbol{\beta}})\}^2K_{h_n}(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}})$$
$$+\frac{1}{n^2}\sum_{i=1}^{n}\{Y_i-\tilde{\mathbf{X}}_i^\top\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}})-(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1(\tilde{\boldsymbol{\beta}})\}^2K_{h_n}(\mathbf{X}_{i1}^\top\tilde{\boldsymbol{\beta}})+\sum_{k=1}^{d}p_{\tilde{\lambda}_k}(|\beta_k|), \tag{6.3}$$

  with respect to $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_d)^\top\in R^d$; denote the minimizer as $\hat{\boldsymbol{\beta}}$.

Go to *Step 1*, and replace $\tilde{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}$ and repeat these two steps until convergence. denote the ultimate estimates as $\hat{\mathbf{a}}=(\hat{a}_0,\cdots,\hat{a}_{d-1})^\top$, $\hat{\mathbf{B}}=(\hat{\mathbf{b}}_1,\cdots,\hat{\mathbf{b}}_n)^\top$, $\hat{\mathbf{F}}=(\hat{\mathbf{f}}_2,\cdots,\hat{\mathbf{f}}_n)^\top=(\hat{\mathbf{f}}_{(0)},\cdots,\hat{\mathbf{f}}_{(d-1)})$ and $\hat{\boldsymbol{\beta}}=(\hat{\beta}_1,\cdots,\hat{\beta}_d)$, respectively. Again, we have suppressed the dependency of these final estimates on $\boldsymbol{\lambda}$. Provided that the initial estimator $\tilde{\boldsymbol{\beta}}$ is close enough to $\boldsymbol{\beta}_0$, the asymptotic properties of these estimators are independent of the choice of $\tilde{\boldsymbol{\beta}}$; see, Theorem 6.2.1 for more details.

6.1 Technical Conditions

Let $\delta_n=(nh_n/\log n)^{-1/2}$, $\tau_n=h_n^2+\delta_n$ and $\boldsymbol{\Theta}_n=\{\boldsymbol{\beta}:|\delta_{\tilde{\boldsymbol{\beta}}}|\le c_1n^{-1/2+c_2}\}$, for some constants $c_1>0$ and $0<c_2<1/10$. Write $\mathscr{T}=\{\mathbf{x}^\top\boldsymbol{\beta}:\mathbf{x}\in\mathscr{D},\boldsymbol{\beta}\in\boldsymbol{\Theta}_n\}$. For any $\boldsymbol{\beta}\in\boldsymbol{\Theta}_n$, let $\delta_\beta=\boldsymbol{\beta}-\boldsymbol{\beta}_0$ and denote by $f(\cdot|\boldsymbol{\beta})$, the probability density of $\mathbf{X}^\top\boldsymbol{\beta}$; for any given $\mathbf{x}\in\mathscr{D}$, write $f(\mathbf{x}^\top\boldsymbol{\beta}|\boldsymbol{\beta})$ as $f(\mathbf{x}|\boldsymbol{\beta})$.

Let $\tilde{\mathbf{X}}_{i(1)}$ and $\tilde{\mathbf{X}}_{i(2)}$ are sub-vectors of $\tilde{\mathbf{X}}_i$ indexed by S_1 and $S_1 \cup S_2$, respectively, i.e. $\tilde{\mathbf{X}}_{i(1)} = (X_{ik}, k \in S_1)$, $\tilde{\mathbf{X}}_{i(2)} = (X_{ik}, k \in S_1 \cup S_2)$. Also for any $\boldsymbol{\beta} \in \boldsymbol{\Theta}_n$, and $t \in \mathscr{T}$, define the following:

$$\Omega(t|\boldsymbol{\beta}) = E(\tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i^\top | \mathbf{X}_i^\top \boldsymbol{\beta} = t),$$

$$\Omega_{11}(t|\boldsymbol{\beta}) = E\{\tilde{\mathbf{X}}_{i(1)} \tilde{\mathbf{X}}_{i(1)}^\top | \mathbf{X}_i^\top \boldsymbol{\beta} = t\},$$

$$\Omega_{22}(t|\boldsymbol{\beta}) = E\{\tilde{\mathbf{X}}_{i(2)} \tilde{\mathbf{X}}_{i(2)}^\top | \mathbf{X}_i^\top \boldsymbol{\beta} = t\},$$

$$\Omega_{20}(t|\boldsymbol{\beta}) = E\{\tilde{\mathbf{X}}_{i(2)} \tilde{\mathbf{X}}_i^\top | \mathbf{X}_i^\top \boldsymbol{\beta} = t\},$$

$$V(t|\boldsymbol{\beta}) = (\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})(t|\boldsymbol{\beta}),$$

$$C(t|\boldsymbol{\beta}) = E\{(\tilde{\mathbf{X}}_{i(1)}^\top \dot{\mathbf{f}}_0(\mathbf{x}))^2 \mathbf{X}_{i\mathbf{x}} \mathbf{X}_{i\mathbf{x}}^\top | \mathbf{X}_i^\top \boldsymbol{\beta} = t\},$$

where all the expectations are taken with respect to $\mathbf{X}_i$ conditional on $\mathbf{X}_i^\top \boldsymbol{\beta} = t$. With a slight abuse of notation, write $\Omega(\mathbf{x}^\top \boldsymbol{\beta}|\boldsymbol{\beta})$ as $\Omega(\mathbf{x}|\boldsymbol{\beta})$, and terms $\Omega_{11}(\mathbf{x}|\boldsymbol{\beta})$, $\Omega_{22}(\mathbf{x}|\boldsymbol{\beta})$, $\Omega_{12}(\mathbf{x}|\boldsymbol{\beta}$, $V(\mathbf{x}|\boldsymbol{\beta})$ and $C(\mathbf{x}|\boldsymbol{\beta})$ should be interpreted in a similar manner. Let $C_0 = E[(f.C)(\mathbf{X}|\boldsymbol{\beta}_0)]$, with $(f.C)(\mathbf{x}|\boldsymbol{\beta}) \equiv f(\mathbf{x}|\boldsymbol{\beta})C(\mathbf{x}|\boldsymbol{\beta})$, and C_{02} denote the $(d - d_0) \times (d - d_0)$ sub-matrix from the lower-right corner of C_0.

The following conditions are assumed throughout the paper unless stated otherwise.

(C1) There exists some constant α, such that for any $k \in S_1$, the second order derivative of function $f_k(.)$, is Hölder continuous with exponent α.

(C2) The density function $f(t|\boldsymbol{\beta})$ is uniformly bounded away from zero in $t \in \mathcal{T}$ and $\boldsymbol{\beta} \in \boldsymbol{\Theta}_n$. Its second order (partial) derivatives are uniformly bounded as functions of $\boldsymbol{\beta} \in \boldsymbol{\Theta}_n$ and $t \in \mathcal{T}$.

(C3) C_{02} is of rank $d - d_0 - 1$ and $\lambda_0(\mathbf{x})$, the smallest eigenvalue of $\Omega(\mathbf{x}|\boldsymbol{\beta}_0)$, is such that

$$\lambda_0 = \min_{\mathbf{x} \in \mathscr{D}} \lambda_0(\mathbf{x}) > 0. \tag{6.4}$$

(C4) $K(.)$ is a symmetric density function with a compact support and a second moment equal to one.

(C5) The smoothing parameter $h_n \propto n^{-1/5}$, while the complexity parameter vector $\boldsymbol{\lambda} = (\lambda_0, \cdots, \lambda_{2d-1}, \tilde{\lambda}_1, \cdots, \tilde{\lambda}_d)^\top$ is chosen such that $|\boldsymbol{\lambda}| \to 0$, as $n \to \infty$.

6.2 Asymptotic properties

Write $\mathbf{f}_0(\mathbf{X}_1) \equiv \mathbf{a}^0 = (a^0_k, \cdots, a^0_{d-1})^\top$, $\mathbf{a}^0_{(2)} = (a^0_k, k \in S_1 \cup S_2)$; for $j = 2, \cdots, n$, $\mathbf{f}^0_j = \mathbf{f}_0(\mathbf{X}_j) - \mathbf{f}_0(\mathbf{X}_1) = (f^0_{jk}, k \in S_0)$, and $\mathbf{f}^0_{j(1)} = (f^0_{jk}, k \in S_1)$; for $j = 1, \cdots, n$, $\mathbf{b}^0_j = \dot{\mathbf{f}}_{n0}(\mathbf{X}_j)$. Correspondingly, we use $\hat{\mathbf{a}}_{(2)}$ and $\hat{\mathbf{f}}^0_{j(1)}$ to denote the estimates of $\mathbf{a}^0_{(2)}$ and $\mathbf{f}^0_{j(1)}$, subvectors of $\hat{\mathbf{a}}$ and $\hat{\mathbf{f}}_j$. Write $\boldsymbol{\beta}_{02} \equiv (\beta_{0,d_0+1}, \cdots, \beta_{0d})^\top$, the vector containing non-zero elements of $\boldsymbol{\beta}_0$, and consider the corresponding

partition of $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^\top, \hat{\boldsymbol{\beta}}_2^\top)^\top$. Further define

$$
\begin{aligned}
&M_0 = E\{(\Omega_{22}.f)(\mathbf{X}|\boldsymbol{\beta}_0)\},\\
&M_{(1)(2)}(\mathbf{x}|\boldsymbol{\beta}) = E[\tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_0(\mathbf{x})\mathbf{X}_{ix(2)}\tilde{X}_{i(1)}^\top|\mathbf{X}_i^\top\boldsymbol{\beta} = \mathbf{x}^\top\boldsymbol{\beta}],\\
&M_{(2)(2)}(\mathbf{x}|\boldsymbol{\beta}) = E[\tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_0(\mathbf{x})\mathbf{X}_{ix(2)}\tilde{X}_{i(2)}^\top|\mathbf{X}_i^\top\boldsymbol{\beta}_0 = \mathbf{x}^\top\boldsymbol{\beta}],\\
&\nu(\mathbf{x}|\boldsymbol{\beta}) = E(\mathbf{X}|\mathbf{X}^\top\boldsymbol{\beta} = \mathbf{x}^\top\boldsymbol{\beta}) - \mathbf{x},
\end{aligned}
$$

where $\mathbf{X}_{ix(2)}$ stands for the subvector of $\mathbf{X}_{ix}$ indexed by $\{d_0 + 1, \cdots, d\}$; $\nu_{(2)}(\mathbf{x}|\boldsymbol{\beta}_0)$ is the subvector of $\nu(\mathbf{x}|\boldsymbol{\beta}_0)$ similarly defined.

We use $a.s.$ to denote almost surely. For an arbitrary index set $\mathcal{Z}$ and a real-valued random matrix $A_n(z)$, we write, $A_n(z) = \mathcal{O}(a_n|\mathcal{Z})$ or $A_n(z) = \mathcal{O}(a_n)$ for simplicity, if

$$
\lim_n \sup_{z\in\mathcal{Z}} |A_n(z)|/a_n = O(1) \quad a.s;
$$

write $A_n(z) = O_p(a_n)$ if $P(\sup_{z\in\mathcal{Z}} |A_n(z)|/a_n = O(1)) \to 1$.

Theorem 6.2.1. *Suppose conditions (C1)-(C5) in Section 6.1 hold and the initial estimator $\tilde{\boldsymbol{\beta}} \in \boldsymbol{\Theta}_n$. In addition, assume the complexity parameter vector $\boldsymbol{\lambda}$ is chosen such that as $n \to \infty$,*

$$
\begin{aligned}
&\frac{\min\{\lambda_k : k \in S_0 \backslash S_1 \cup S_2\}}{\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}|} \to \infty;\\
&\frac{\min\{\lambda_{k+d} :\ k \in S_0 \backslash S_1\}}{\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}|} \to \infty,\\
&\frac{\min\{\tilde{\lambda}_k,\ 1 \le k \le d_0\}\}}{(\log n/n)^{1/2} + |\delta_{\tilde{\boldsymbol{\beta}}}|} \to \infty.
\end{aligned}
$$

Then we have

(a) [Sparsity] $\Pr(\max_{k \notin S_1 \cup S_2} |\hat{a}_k| = 0,$ *for large enough* $n) = 1;$

$\Pr(\max_{k \notin S_1} |\hat{\mathbf{f}}_{(k)}| = \mathbf{0},$ *for large enough* $n) = 1;$

$\Pr(\max_{1 \le k \le d_0} |\hat{\beta}_k| = 0,$ *for large enough* $n) = 1;$

(b) As $n \to \infty,$

$$\begin{aligned}
\hat{\mathbf{a}}_{(2)} - \mathbf{a}^0_{(2)} &= M_0^{-1} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta}_0) \, [\tilde{X}_{i(2)} - \Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta}_0) \\
&\quad \times \{\Omega_{11}(\mathbf{X}_i|\boldsymbol{\beta}_0)\}^{-1} \tilde{X}_{i(1)}] + o_p(n^{-1/2}|\mathscr{D}, \Theta_n), \\
\hat{\mathbf{f}}_{j(1)} - \mathbf{f}^0_{j(1)} &= [f.\Omega_{11}]^{-1}(\mathbf{X}_j|\boldsymbol{\beta}_0) \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \tilde{X}_{i(1)} K_{h_n}(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_0) \\
&\quad + \frac{1}{2} h_n^2 \ddot{\mathbf{f}}_0(\mathbf{X}_j) + \mathscr{O}(h_n \tau_n|\mathscr{D}, \Theta_n), \\
\hat{\mathbf{b}}_j - \mathbf{b}^0_j &= [\Omega(\mathbf{X}_j|\boldsymbol{\beta}_0)]^{-1} \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_0)(\mathbf{X}_{ij}^\top \boldsymbol{\beta}_0 / h_n) \tilde{\mathbf{X}}_i \varepsilon_i \\
&\quad + \mathscr{O}(h_n \tau_n|\mathscr{D}, \Theta_n).
\end{aligned}$$

(c) With C_{02}^+ *being the Moore-Penrose inverse of* C_{02}*,*

$$\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{02} = \frac{2}{n} C_{02}^+ \sum_i \varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta}) M(\mathbf{X}_i) + \mathscr{O}(h_n \tau_n),$$

where

$$\begin{aligned}
M(\mathbf{X}_i) = \nu_{(2)}(\mathbf{X}_i|\boldsymbol{\beta}_0) \tilde{\mathbf{X}}_i^\top \ddot{\mathbf{f}}_0(\mathbf{X}_i) - (M_{(1)(2)} \Omega_{11}^{-1})(\mathbf{X}_i|\boldsymbol{\beta}_0) \tilde{\mathbf{X}}_{i(1)} \\
- E\{(f.M_{3(2)}^\top)(\mathbf{X}|\boldsymbol{\beta}_0)\} M_0^{-1} [\tilde{X}_{i(2)} - (\Omega_{21} \Omega_{11}^{-1}) \\
\times (\mathbf{X}_i|\boldsymbol{\beta}_0) \tilde{X}_{i(1)}].
\end{aligned}$$

CHAPTER 7

SIMULATION STUDY

In this section, we use simulation studies to demonstrate and augment our theoretical results and to evaluate the accuracy of the proposed model selection and estimation procedure. The kernel function we used in this section is Epanechnikov kernel $K(t) = 0.75(1-t^2)_+$. Following the idea in Section 5.1.2, we select the bandwidth in terms of $H = 0.6(d_n/n)^{0.2}$. The tuning parameters are determined by the criterion described in Section 5.2.

We initially show that the goodness of the ultimate estimators from our proposed approach is independent of the choice of initial value $\tilde{\boldsymbol{\beta}}$ in Section 4.2.2. Then present several simulated examples to assess the accuracy of the proposed model selection and estimation procedure and also examine the oracle property of the proposed estimators in Section 7.2.

7.1 Sensitivity to the choice of initial value $\tilde{\boldsymbol{\beta}}$

In this section, we consider the same simulation model (5.1) in Section 5.1. Based on the constraints that $||\boldsymbol{\beta}|| = 1$, $\beta_1 > 0$ and $\beta_{d_n} \neq 0$, we provide several candidate initial values $\tilde{\boldsymbol{\beta}}$ as follows:

$$\tilde{\boldsymbol{\beta}}_{(1)} = \frac{1}{\sqrt{d}}(1,1,1,\ldots,1)^\top, \qquad \tilde{\boldsymbol{\beta}}_{(2)} = \frac{1}{\sqrt{3}}(1,1,0,\ldots,1)^\top,$$
$$\tilde{\boldsymbol{\beta}}_{(3)} = \frac{1}{\sqrt{2}}(1,0,0\ldots,1)^\top, \qquad \tilde{\boldsymbol{\beta}}_{(4)} = \frac{1}{\sqrt{3}}(2,0,0,\ldots,1)^\top.$$

We conduct simulation on the regression example (5.1) with dimension $d = 7$ in 1000 replications, each with the sample size $n = 600$.

Start with each candidate initial value given above; our iterative procedure will be executed to obtain the corresponding Correct Rate (defined in Section 5.2), MSE of the estimated index parameters and RMISE of the estimated functional coefficients (defined in Section 5.1), respectively. The bandwidth is chosen as $H = 0.30$ and the tuning parameters are chosen to be $\lambda_f = 4, \lambda_\beta = 40$. We summarise the simulated results in Table 7.1.

Table 7.1: Sensitivity to the choice of initial value $\tilde{\boldsymbol{\beta}}$ on the regression model with dimension $d = 7$

Initial value	Correct Rate	MSE	RMISE
$\tilde{\boldsymbol{\beta}}_{(1)}$	0.989	0.003606	0.16129
$\tilde{\boldsymbol{\beta}}_{(2)}$	0.989	0.003599	0.16217
$\tilde{\boldsymbol{\beta}}_{(3)}$	0.988	0.003592	0.16182
$\tilde{\boldsymbol{\beta}}_{(4)}$	0.987	0.003604	0.16289

From the statistic of each criterion in Table 7.1, we can easily find that the diverse candidates of initial value lead to almost the same selection and estimation results, which all performs remarkably well.

Furthermore, by setting the dimension of the regression example (5.1) $d = 20$, we will conduct another simulation to verify the fact again in the high-dimensional situation. The simulated results are reported in Table 7.2.

Table 7.2: Sensitivity to the choice of initial value $\tilde{\boldsymbol{\beta}}$ on the regression model with dimension $d = 20$

Initial value	Correct Rate	MSE	RMISE
$\tilde{\boldsymbol{\beta}}_{(1)}$	0.946	0.004673	0.14091
$\tilde{\boldsymbol{\beta}}_{(2)}$	0.945	0.004632	0.14106
$\tilde{\boldsymbol{\beta}}_{(3)}$	0.949	0.004571	0.14012
$\tilde{\boldsymbol{\beta}}_{(4)}$	0.950	0.004528	0.13809

From the statistic results in Table 7.1 and Table 7.2, we conclude that the our proposed approach of model selection and estimation is not sensitive to the choice of initial value $\tilde{\boldsymbol{\beta}}$.

7.2 Simulation examples

In this section, we shall begin with two varying coefficients examples in a modest dimension that is $d = 7$. Then, to further illustrate the goodness of the proposed method in higher dimensionality, we will increase the dimension of both simulated examples to $d = 20$. Meanwhile, we will compare the performance of the proposed method in different dimensions on model selection, structure specification and the accuracy of estimation.

Similar to the regression example (5.1), we consider the following two examples of varying coefficient models.

1. $Y_i = 2\exp(-Z_i^2) + X_{i2} + \varepsilon_i$, with $Z_i = \mathbf{X}_i^\top \boldsymbol{\beta}_0 = \frac{1}{3}(2X_{i1} + 2X_{i2} + X_{id})$,

2. $Y_i = 2\cos(0.5\pi Z_i) + Z_i X_{i1} + 4\exp(-Z_i^2) X_{i2} + \varepsilon_i$, with $Z_i = \mathbf{X}_i^\top \boldsymbol{\beta}_0 = \frac{1}{3}(X_{i1} + 2X_{i2} + 2X_{id})$,

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^\top$, for $i = 1, \ldots, n$, are normally distributed independent random vectors and noise ε_i are

independent N$(0,1)$ random variables. The regression models are based on the form (4.2) with $\boldsymbol{\beta}_0 = \frac{1}{3}(2,2,0,\ldots,1)^\top$ and $\boldsymbol{\beta}_0 = \frac{1}{3}(1,2,0,\ldots,2)^\top$, respectively. we will firstly consider these models in the dimension of $d = 7$. Then, another simulation will be conducted on the same regression functions but the dimensions of models is replaced by $d = 20$. For each case, we conduct simulation with sample size $n = 600$, in a total of 1000 replications.

To evaluate the performance of model selection, we report the ratio of correct, under-fitted, over-fitted and other models. Whenever the resulting model simultaneously detects the true model and identifies the modeling structure correctly, we classify it as a "correct model". Whenever the estimated model eliminates at least one significant covariates but does not include any irrelevant covariates, we classify it as an "under-fitted model". Whenever the estimated model includes at least one insignificant covariates but does not miss any relevant covariates, it is labelled as an "over-fitted model". The "other models" means that the estimated model not only includes the irrelevant covariates but also ignores relevant covariates.

The simulation results are reported in Table 7.3. We can notice that, in all cases, the percentage of the correctly selected models is no less than 94%, which verifies that the proposed method indeed select the true model consistently.

Besides, the fact that the ratio of correctly fitted models increases slightly as the dimension decrease also makes sense.

Table 7.3: The ratios of model selection in 1000 replications

d	Correct	Under-fitted	Over-fitted	Others
		Example 1		
7	0.972	0.013	0.015	0
20	0.943	0.032	0.023	0.002
		Example 2		
7	0.971	0.007	0.022	0
20	0.940	0.023	0.037	0

Apart from assessing the correctness of the selection, we will also evaluate the estimation accuracy of the proposed estimate. In particular, instead of computing the MSE of the vector of estimated index parameter, in this section, we calculate the MSE in component-wise manner, namely, compute the MSE of the estimate with respect to each significant index parameter, which can be defined as

$$\mathrm{MSE}_{\beta_k} = \frac{1}{L}\sum_{l=1}^{L}(\hat{\beta}_k^l - \beta_k)^2,$$

where $\hat{\beta}_k^{(l)}, k = 1,2,\ldots,d_n, l = 1,2,\ldots,L$, is the estimate from the $l-$th replication with respect to the $k-$th index parameter in the $l-$th; we also figure out the RMISE for the estimates of each relevant coefficient, which is approximated as follows

$$\text{Relative MISE}_{f_k(\cdot)} \approx \frac{1}{L}\sum_{l=1}^{L}[\frac{\sum_{j=1}^{n}(\hat{f}_k^{(l)}(z_j)-f_k(z_j))^2}{\sum_{j=1}^{n}f_k(z_j)^2}];$$

where $\hat{f}_k^{(l)}(\cdot), k = 0,1,\ldots,d_n-1, l = 1,2,\ldots,L$, is estimator of the $k-$th functional coefficient in the $l-$th replication and $z_j = x_j^T\hat{\boldsymbol{\beta}}\, j = 1,\ldots,n.$ with some estimator $\hat{\boldsymbol{\beta}}$.

Additionally, introducing a benchmark to compare with is essential for evaluating the accuracy of the estimation. We employ the "oracle estimators" as the benchmark, who are the estimators of coefficients from the models that have already been correctly selected as the true model and whose estimation procedure is free from penalised approaches. Hence, we will report the RMISE and MSE of oracle estimators as well. The simulation is also conducted in 1000 replications whose results are summarized in Table 7.4.

As we can see from Table 7.4, all the values of MSE and RMISE are reasonably small, who gradually become smaller with the decrease of the dimension of models. Besides, the oracle estimators are always more accurate than the estimators from other models. Both of these findings confirm the accuracy of our estimation. Therefore, we conclude that our proposed method can simultaneously select the true model correctly and estimate the model precisely.

Inspired by the fact that our proposed selection method can

consistently select the true model and the oracle estimators from the true model outperform the penalised estimates, we decide to improve our proposed approach by the following procedures:

1. Apply the model selection method proposed in Section 4.2.2 to select a sub-model which is expected to be the true model;

2. Estimate the selected sub-model by the penalty free iterative approach proposed in Section 3.2.2 to obtain the final estimates for index parameters and functional coefficients, respectively.

This modified procedure will be used in the real data analysis in Chapter 8.

Table 7.4: The RMISEs and MSEs of the varying and constant parameters

Model 1			
	$d=7$	$d=20$	Oracle
MSE_{β_1}	0.0357	0.0452	0.0248
MSE_{β_2}	0.0283	0.0299	0.0232
MSE_{β_3}	0.0105	0.0111	0.0048
MSE_{c_2}	0.0067	0.0523	0.0005
$RMISE_{f_0}$	0.1584	0.1880	0.0699

Model 2			
	$d=7$	$d=20$	Oracle
MSE_{β_1}	0.0020	0.0119	0.0008
MSE_{β_2}	0.0015	0.0031	0.0005
MSE_{β_3}	0.0016	0.0046	0.0006
$RMISE_{f_0}$	0.2017	0.2136	0.1556
$RMISE_{f_1}$	0.0535	0.2424	0.0386
$RMISE_{f_2}$	0.0766	0.0964	0.0630

NOTE: The rows labeled by $d=7$ and $d=20$ represent the estimators from the proposed method in the 7-dimensional models and 20-dimensional models, respectively. the rows labeled by “oracle” depict the oracle estimators. In Model 1, there exists a constant coefficient, which is denoted by c_2 and the column labeled as $MSE_{c_2}$is the MSE of this costant coefficient.

CHAPTER 8

REAL DATA ANALYSIS

In this chapter, we are going to illustrate the Iterative kernel smoothly clipped absolute deviation penalty method by two real data examples. We first construct the single index varying coefficient model on both datasets to solve regression problems and then use the purposed method to select and fit the models.

8.1 Real data example I

We consider here an environmental data set from Hong Kong, which was collected from January 1, 1994, to December 31, 1995 (courtesy of Professor T. S. Lau). In the dataset, we take the numbers of daily total hospital admissions for circulatory and respiratory problems as the response and the following covariates as the X-variables: SO_2 (coded by x_1), NO_2 (coded

by x_2), dust (coded by x_3), ozone (coded by x_4), temperature (coded by x_5), the change in temperature (which is the absolute value of the temperature difference between two time points, coded by x_6) and humidity (coded by x_7).

Among all these environmental factors, we would like to detect which factors are significantly relevant to the number of daily total hospital admissions for circulatory and respiratory problems (whose logarithm is coded by y), and whether the effect of those factors vary over a comprehensive environment index (coded by $z = \mathbf{x}^\top \boldsymbol{\beta}, \mathbf{x} = (x_1, \ldots, x_7)^\top$), which is a linear combination of index parameter (coded by $\boldsymbol{\beta}$) and some of the collected environmental factors. Before the modelling, as the variables are in different units, we need to standardize the data such that they have sample mean 0 and sample covariance matrix I_d. To realize the objective, we employ a single index varying coefficient model as follows:

$$y_i = f_0(z_i) + f_1(z_i)x_{i1} + f_2(z_i)x_{i2} + \ldots + f_6(z_i)x_{i6} + \varepsilon_i,$$
$$\text{with } z_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_7 x_{i7}, \quad (8.1)$$

where we take $x_{i0} = 1$ as the intercept term. We apply the proposed model selection method to identify the sub-model with significant variables and important index parameters. The tuning parameters are selected by the GIC approach described in Section 5.2.

The selected results suggest that SO_2 , NO_2, temperature and the change in temperature have significant effects on the response and all of their coefficients are functional coefficients. Meanwhile, the important index parameters are selected as $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_5, \beta_7)^\top$, which indicates that the comprehensive environment index includes SO_2 , NO_2, temperature and humidity. Hence, we have such a selected model

$$\begin{gathered} y_i = f_0(z_i) + f_1(z_i)x_{i1} + f_2(z_i)x_{i2} + f_5(z_i)x_{i5} + f_6(z_i)x_{i6} + \varepsilon_i, \\ \text{with } z_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_5 x_{i5} + \beta_7 x_{i7}. \end{gathered} \tag{8.2}$$

It states in Chapter 7.2 that, if the correct sub-model has been consistently selected, smoothing the sub-model via the penalty-free iterative approach leads to the estimates can be treated as the approximation of "oracle estimators". The "oracle estimators" are shown to outperform the penalised estimates. Accordingly, We fit the selected model (8.2) by the proposed penalty-free iterative approach, and hence work out the estimators of the varying coefficients with respect to the significant covariates and the estimators of index parameters, which are

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_5, \hat{\beta}_7)^\top = (0.5981, 0.7226, 0.1177, 0.3260)^\top.$$

However, from an intuitive assessment, we notice that the estimated curves are not very smooth. To ameliorate this

issue, we decide to employ a "two-step" local linear regression as a modification to produce better-estimated cures and, more important, to get a more accurate estimation result. Precisely, this straightforward smoothing is implemented as follows.

Specify a set $S_0 = \{0, 1, \ldots, d-1\}$ and its subset $S_1 = \{k : k \in S_0,\ \hat{f}_k(\cdot) \text{ is not constant}\}$. From the aforementioned procedure, we have obtained the local estimators of $f_k(\cdot)$, $k \in S_1$, which is denoted by $\hat{\mathbf{f}}_k = (\hat{f}_{1k}, \ldots, \hat{f}_{nk})^\top \in \mathbb{R}^n$, $k \in S_1$ and the estimator of index parameters $\hat{\beta}_k$, $k \in \{k : k = 1, \ldots, d,\ \hat{\beta}_k \neq 0\}$. We consider $(\hat{f}_{ik}, \hat{z}_i)$, $i = 1, \ldots, n$ as the observations. A regression problem assumes that

$$\hat{f}_{ik} = m_k(\hat{z}_i) + \varepsilon_i,$$

where $\hat{z}_i = \mathbf{x_i}^\top \hat{\boldsymbol{\beta}}$ and ε_i is the random noise. The local linear estimators for the unknown function $m_k(\hat{z}_j)$, $j = 1, \ldots, n$ is obtained by minimizing the sum

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \hat{f}_{ik} - [A_{jk} + B_{jk}(\hat{z}_i - \hat{z}_j)] \right\}^2 K_h(\hat{z}_i - \hat{z}_j), \tag{8.3}$$

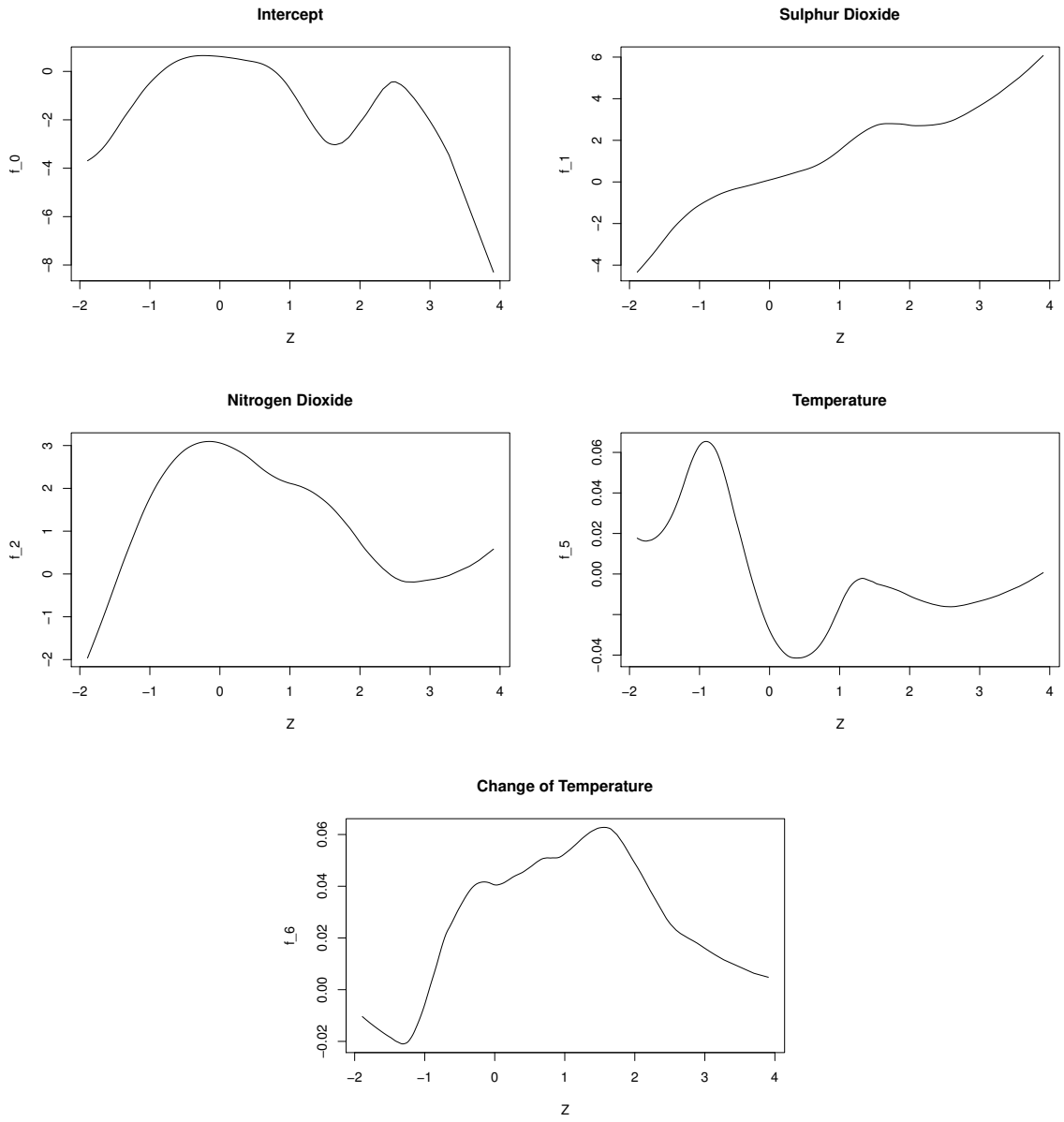
with respect to A_{jk} and B_{jk}. We define the minimiser of (8.3) as $\hat{A}_{jk}$, which is the estimator of $m_k(\hat{z}_j)$. Consequently, the fitted value of $\hat{f}_{jk}$, $j = 1, \ldots, n$ is given by $\hat{A}_{jk}$. By solving the minimisation problem (8.3), we obtain the resulting estimators of $f_k(\cdot)$ from the "two-step" method as $\hat{\mathbf{A}}_{jk} = (\hat{A}_{1k}, \ldots, \hat{A}_{1k})$.

Furthermore, in order to verify that the "two-step" local linear regression leads to better estimates, we next compare

the performance of the corresponding estimators from our purposed methods with those from the "two-step" methods. The "Leave-p-out Cross-Validation (LpO CV)" is introduced as the measurement. The "LpO CV" of the "two-step" methods is 0.0316, which is slightly smaller than the "LpO CV" of our purposed method, which is 0.0375. Hence, implementing the "two-step" local linear regression to smooth the curves is reasonable. We now visualise our ultimate estimation results in Figure 8.1.

As illustrated in Figure 8.1, the coefficients of those three factors are unlikely to be null or other constants, and they all vary over the range of comprehensive environment index. Besides, we can easily discover some simple but convincing conclusions from the estimated results. Firstly, from the estimated curves, we find that both NO_2 and SO_2 become more damaging to people's circulatory or respiratory system in the warm and moist climate. Secondly, the main air pollution puts people at more risk for sickness with the increasing of their concentration. Meanwhile, when the concentration of the toxic gas is above some certain level (when the index z is larger than zero), both NO_2 and SO_2 always have a positive impact on people's circulatory or respiratory problems. In addition, from the last estimated curve in Figure 8.1, we notice that the change of temperature has an index-varying positive effect on the daily number of total hospital admissions when the

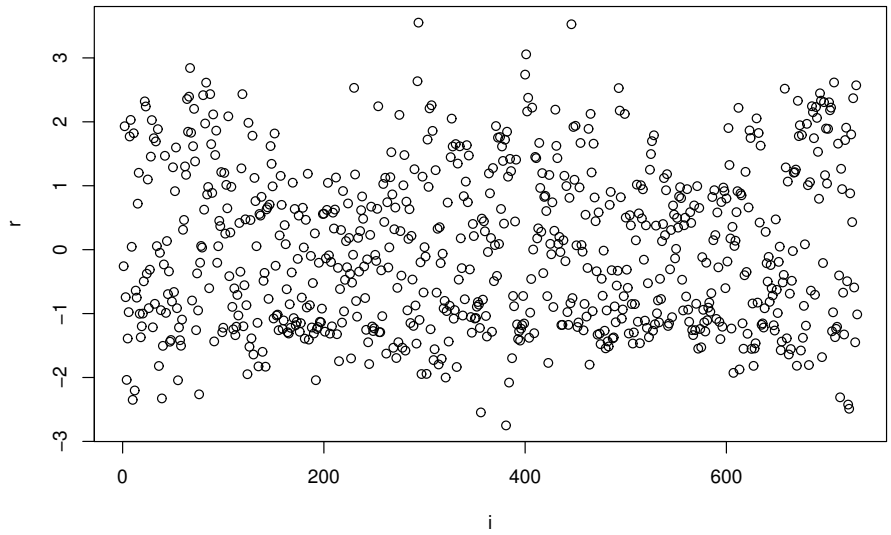
Figure 8.1: Estimated curves of varying coefficients in the selected model



index is positive. This insight is in line with the fact that large fluctuations in temperature may exacerbate people's health conditions and trigger many kinds of pains and diseases.

Moreover, we would like to analyse our estimated results by evaluating the residuals between y_i (the logarithm of the number of daily total hospital admission) and its estimator $\hat{y}_i$. We depict the residuals in figure 8.2.

Figure 8.2: Residuals



It can be seen in Figure 8.2, there is no obvious tendency, which also corroborates our purposed selection and estimation methods decently.

8.2 Real data example II

We now illustrate the application of the proposed methodology in Boston housing data, which has been analysed in numerous amounts of literature includes Fan and Huang (2005) and

Wang and Xia (2009). The data set consists of 506 US boroughs in the Boston area. The response variable is the median value of owner-occupied homes (MEDV) in 1970 and there are thirteen factors can be taken into account, some of which may affect the variation in housing value significantly. The description of these thirteen factors serving as the covariates are given in Table 8.1:

Table 8.1: The covariates in Boston houing dataset

CRIM (x_1)	per capita crime rate by town
ZN (x_2)	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS (x_3)	proportion of non-retail business acres per town
CHAS (x_4)	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX (x_5)	nitric oxides concentration (parts per 10 million)
RM (x_6)	average number of rooms per dwelling
AGE (x_7)	proportion of owner-occupied units built prior to 1940
DIS (x_8)	weighted distances to five Boston employment centres
RAD (x_9)	index of accessibility to radial highways
TAX (x_{10})	full-value property-tax rate per \$10,000
PTRATIO (x_{11})	pupil-teacher ratio by town
B (x_{12})	$1000(Bk-0.63)^2$ where Bk is the proportion of blacks by town
LSTAT (x_{13})	percentage of lower status of the population

In this empirical analysis, we are primarily interested in the following three aspects.

1. Identifying which factors among the all the collected

factors contribute significantly to MEDV.

2. Revealing whether the impacts of the relevant factors are constant or vary over an index variable.

3. Detecting which factors are the real components of the index.

To fulfil the objectives, we start by considering a single index varying coefficient model

$$\begin{aligned} y_i = f_0(z_i) + f_1(z_i)x_{i1} + f_2(z_i)x_{i2} + \ldots + f_{12}(z_i)x_{i12} + \varepsilon_i, \\ \text{with}\, z_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{13} x_{i13}, \end{aligned} \tag{8.4}$$

where $x_{i0} = 1, i = 1, \ldots, 50$ is set as the intercept term. To unify the scale of each covariate, we standardize the covariates and response firstly. Then, we apply the proposed method to select the true model in (8.4).

The estimated results suggest that CRIM, ZN, CHAS, NOX, AGE, RAD, TAX, PTRATIO are the significant covariates to the response. Among these relevant factors, only NOX and AGE have a constant impact on MEDV, the rest of them affect the median value of owner-occupied homes in varying significant level. In addition, we identify that the variable CRIM is the most crucial component of the index, which also includes covariates ZN and LSTAT. Therefore, we have such a selected model

$$\begin{aligned} y_i =& f_0(z_i) + f_1(z_i)x_{i1} + f_2(z_i)x_{i2} + f_4(z_i)x_{i4} + C_5(z_i)x_{i5} \\ &+ C_7(z_i)x_{i7} + f_9(z_i)x_{i9} + f_{10}(z_i)x_{i10} + f_{11}(z_i)x_{i11} + \varepsilon_i, \\ \text{with}\, z_i =& \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{13} x_{i13}, \end{aligned}$$

where C_5 and C_7 present the constant coefficient of covariate NOX and AGE, respectively.

We next introduce a similar estimation procedure without penalised approaches to estimate the specified model. By applying the proposed method, we obtain the estimators of index parameters, which are

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_{13})^\top = (0.9453, 0.2861, 0.1567)^\top,$$

the estimated constant coefficients

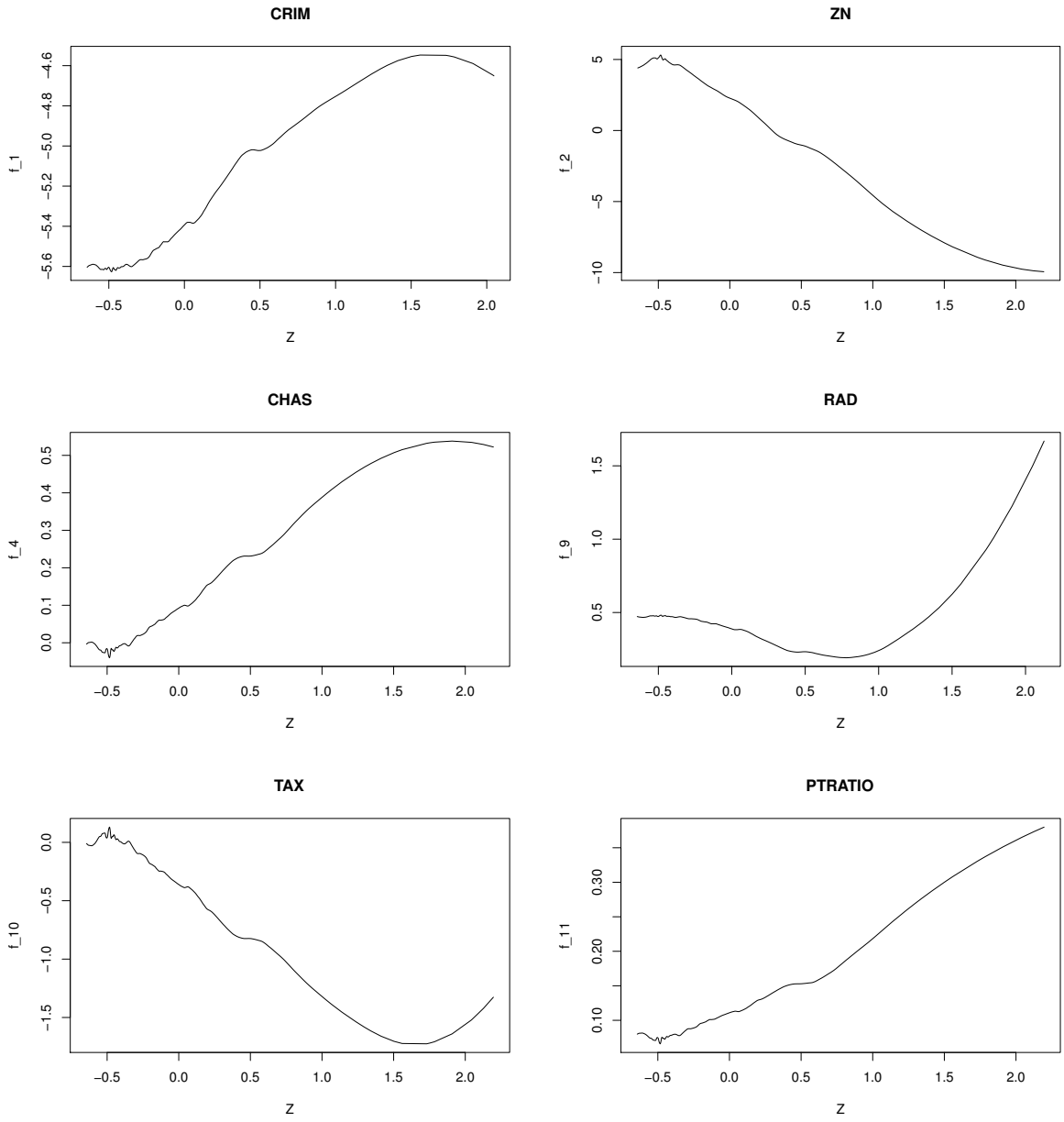
$$\hat{C}_5 = 0.3752, \quad \hat{C}_7 = 0.0808,$$

and the estimated curves of all the varying coefficients of the corresponding covariate, which are provided in Figure 8.3.

Apparently, all the curves in Figure 8.3 indicate that those six coefficients are unlikely to be null or other constants. More importantly, we would like to explore some more insights from the estimated results.

We can see visually that the estimated functional coefficient of CRIM is always negative, which reflects the fact that crimes have a significant negative impact on prices of a house.

Figure 8.3: Estimated curves of the varying coefficients

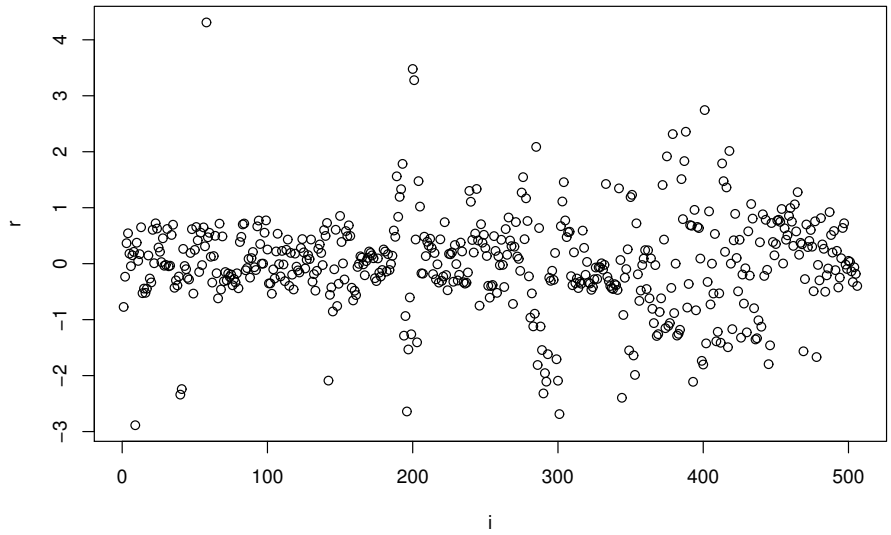


Besides, as the main component of the index is CRIM, we can state from the estimates that the negative impacts of crimes are more sensitive to property value when the CRIM stays at a very low level and the marginal impact of crime

decreases. From the plot of the coefficient of ZN, we find that the large property is highly demanded in the area with good public security but largely devalued in those unsafe areas. Meanwhile, the fact that located beside the river, good traffic facilities and adequate education resource all have positive effects on the property value also makes sense. Also, it can be visually found that the effect of property taxes on house values is consistently adverse once the index is positive. Additionally, houses based in the community with high crime rate are more impressionable to the property taxes.

Following the idea in section 8.1, we also report the estimated results in figure 8.4 for further evaluation.

Figure 8.4: Residuals



As it is shown in Figure 8.4, there is no obvious tendency,

which also corroborates that the purposed selection and estimation methods is quite substantial.

CHAPTER 9

PROOF OF THEORETICAL RESULTS

In Chapter 9.1, we give some technical lemmas which are needed to prove the asymptotic theory in Chapter 6 and their proofs. Then, we provide the proofs of the main theoretical results in Chapter 9.2. Like Chapter 6, this chapter based on my submitted paper "An Iterative Approach for Model Selection in Single-index Varying Coefficient Models", which is the joint work with Prof. Efang Kong and Prof. Wenyang Zhang.

9.1 Lemmas and Proofs

Under (C2), (C5) and (C6) in Section 6.1, the following results are quite standard in the literature on strong uniform convergence for nonparametric smoothing, e.g. Masry (1996) and

Pollard (1984). implies

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top(\mathbf{X}_{i\mathbf{x}}^\top\boldsymbol{\beta}/h_n)K_{h_n}(\mathbf{X}_{i\mathbf{x}}^\top\boldsymbol{\beta})\\
&= h_n[f.\Omega]'(\mathbf{x}|\boldsymbol{\beta}) + \mathscr{O}(|\delta_{\boldsymbol{\beta}}|\tau_n|\mathscr{D},\ \boldsymbol{\Theta}_n)\\
&= h_n[f.\Omega]'(\mathbf{x}|\boldsymbol{\beta}_0) + \mathscr{O}(|\delta_{\boldsymbol{\beta}}| + \tau_n|\mathscr{D},\ \boldsymbol{\Theta}_n), \qquad (9.1)
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top(\mathbf{X}_{i\mathbf{x}}^\top\boldsymbol{\beta}/h_n)^l K_{h_n}(\mathbf{X}_{i\mathbf{x}}^\top\boldsymbol{\beta})\\
&= [f.\Omega](\mathbf{x}|\boldsymbol{\beta}) + \mathscr{O}(\tau_n|\mathscr{D},\ \boldsymbol{\Theta}_n) \qquad (l=0,2)\\
&= [f.\Omega](\mathbf{x}|\boldsymbol{\beta}_0) + \mathscr{O}(|\delta_{\boldsymbol{\beta}}| + \tau_n|\mathscr{D},\ \boldsymbol{\Theta}_n), \qquad (9.2)
\end{aligned}$$

where $[f.\Omega](\mathbf{x}|\boldsymbol{\beta}) = f(\mathbf{x}|\boldsymbol{\beta})\Omega(\mathbf{x}|\boldsymbol{\beta})$, and $[f.\Omega]'(\mathbf{x}|\boldsymbol{\beta})$ denotes the matrix of element-wise (first order) derivative of $[f.\Omega](t|\boldsymbol{\beta})$ with respect to t, and evaluated at $t = \mathbf{x}^\top\boldsymbol{\beta}$.

We first establish the asymptotic properties relating to $\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}}) = (\hat{a}_k(\tilde{\boldsymbol{\beta}}))$, $\hat{\mathbf{F}}(\tilde{\boldsymbol{\beta}}) := (\hat{f}_{jk}(\tilde{\boldsymbol{\beta}})) = (\hat{\mathbf{f}}_2(\tilde{\boldsymbol{\beta}}),\cdots,\hat{\mathbf{f}}_n(\tilde{\boldsymbol{\beta}}))^\top$, and $\hat{\mathbf{B}}(\tilde{\boldsymbol{\beta}}) := (\hat{b}_{jk}(\tilde{\boldsymbol{\beta}})) = (\mathbf{b}_1(\tilde{\boldsymbol{\beta}}),\cdots,\mathbf{b}_n(\tilde{\boldsymbol{\beta}}))^\top$, the minima of (6.2) with initial estimate $\tilde{\boldsymbol{\beta}}$, as estimates of $\hat{\mathbf{a}}^0$, $\mathbf{F}^0 := (f_{jk}^0) = (\mathbf{f}_2^0,\cdots,\mathbf{f}_n^0)^\top = (\mathbf{f}_{(0)}^0,\cdots,\mathbf{f}_{(d-1)}^0)$, and $\mathbf{B}^0 := (b_{jk}^0) = (\mathbf{b}_1^0,\cdots,\mathbf{b}_n^0)^\top$, respectively. For any real matrix A, let $|A|_\infty$ denote the greatest among the absolute values of its elements.

Lemma 9.1.1. *Under conditions in Theorem 6.2.1, we have*

$$|\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}}) - \mathbf{a}^0|_\infty = \mathscr{O}(|\delta_{\tilde{\boldsymbol{\beta}}}| + \tau_n|\boldsymbol{\Theta}_n), \quad |\hat{\mathbf{F}}(\tilde{\boldsymbol{\beta}}) - \mathbf{F}^0|_\infty = \mathscr{O}(|\delta_{\tilde{\boldsymbol{\beta}}}| + \tau_n|\boldsymbol{\Theta}_n),$$
$$|\hat{\mathbf{B}}(\tilde{\boldsymbol{\beta}}) - \mathbf{B}^0|_\infty = \mathscr{O}(\tau_n|\boldsymbol{\Theta}_n).$$

Proof of Lemma 9.1.1 Write $\alpha_n = \tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}|$. It thus suffices to show that with probability one, for large enough $C > 0$ and n,

$$\min_{\boldsymbol{\beta}\in\boldsymbol{\Theta}_n} \{Q_n(\mathbf{a}^0 + C\alpha_n\mathbf{w}, \mathbf{F}^0 + C\alpha_n\mathbf{U}, \mathbf{B}^0 + C\tau_n\mathbf{V}|\boldsymbol{\beta}, \boldsymbol{\lambda}) - Q_n(\mathbf{a}^0, \mathbf{F}^0, \mathbf{B}^0|\boldsymbol{\beta}, \boldsymbol{\lambda})\} > 0, \tag{9.3}$$

for any generic vector $\mathbf{w} = (w_k)_{0\le k\le d-1}$, matrices $\mathbf{U} = (\mathbf{u}_2, \cdots, \mathbf{u}_n)^\top = (\mathbf{u}_{(0)}, \cdots, \mathbf{u}_{(d-1)}) \in R^{(n-1)\times d}$ and $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_n)^\top = (\mathbf{v}_{(0)}, \cdots, \mathbf{v}_{(d-1)}) \in R^{n\times d}$, such that $|\mathbf{w}|_\infty = 1$, $|\mathbf{U}|_\infty = 1$, $|\mathbf{V}|_\infty = 1$. We first show that with probability one, for large enough $C > 0$ and n,

$$\min_{\boldsymbol{\beta}\in\boldsymbol{\Theta}_n} \left[\bar{Q}_n(\mathbf{a}^0 + C\alpha_n\mathbf{w}, \mathbf{F}^0 + C\alpha_n\mathbf{U}, \mathbf{B}^0 + C\tau_n\mathbf{V}|\boldsymbol{\beta}) - \bar{Q}_n(\mathbf{a}^0, \mathbf{F}^0, \mathbf{B}^0|\boldsymbol{\beta})\right] \ge 0, \tag{9.4}$$

where the equality holds if and only if $\mathbf{w} = \mathbf{0}$, $\mathbf{U} = \mathbf{0}$ and $\mathbf{V} = \mathbf{0}$, and

$$\begin{aligned}\bar{Q}_n(\mathbf{a}, \mathbf{F}, \mathbf{B}|\boldsymbol{\beta}) &= \frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a} + \mathbf{f}_j) - (\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_j\}^2 \\ &\quad \times K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) \\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top\mathbf{a} - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1\}^2 K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta});\end{aligned}$$

specifically, for any large enough C and n, with $\mathbf{f}_1^0 \equiv \mathbf{0}$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a}^0 + C\alpha_n\mathbf{w} + \mathbf{f}_j^0 + C\alpha_n\mathbf{u}_j) \\
&-(\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top(\mathbf{b}_j^0 + C\tau_n\mathbf{v}_j)\}^2 K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) \\
&\geq \frac{1}{2}C^2\alpha_n^2(\mathbf{w}+\mathbf{u}_j)^\top[f.\Omega](\mathbf{X}_j|\boldsymbol{\beta})(\mathbf{w}+\mathbf{u}_j) \\
&+\frac{1}{2}C^2\tau_n^2{\mathbf{v}_j}^\top[f.\Omega](\mathbf{X}_1|\boldsymbol{\beta})\mathbf{v}_j,
\end{aligned} \tag{9.5}$$

uniformly in $j = 1, \cdots, n$, and $\boldsymbol{\beta} \in \Theta_n$. To illustrate, consider

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a}^0 + C\alpha_n\mathbf{w}) - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top(\mathbf{b}_1^0 + C\tau_n\mathbf{v}_1)\}^2 \\
&\times K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta}) - \frac{1}{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top\mathbf{a}^0 - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1^0\}^2 K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta}) \\
&= C^2\alpha_n^2\mathbf{w}^\top\Big[\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\Big]\mathbf{w} \\
&+C^2\tau_n^2\mathbf{v}_1^\top\Big[\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)^2 K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\Big]\mathbf{v}_1 \\
&+2C^2\alpha_n\tau_n\mathbf{w}^\top\Big[\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\Big]\mathbf{v}_1 \\
&-2C\alpha_n\mathbf{w}^\top\Big[\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\{Y_i - \tilde{\mathbf{X}}_i^\top\mathbf{a}^0 - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1^0\}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\Big] \\
&-2C\tau_n\mathbf{v}_1^\top\Big[\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\{Y_i - \tilde{\mathbf{X}}_i^\top\mathbf{a}^0 - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1^0\} \\
&\times K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\Big] \\
&\geq \frac{1}{2}C^2\alpha_n^2\mathbf{w}^\top[f.\Omega](\mathbf{X}_1|\boldsymbol{\beta})\mathbf{w} + \frac{1}{2}C^2\tau_n^2{\mathbf{v}_1}^\top[f.\Omega](\mathbf{X}_1|\boldsymbol{\beta})\mathbf{v}_1,
\end{aligned} \tag{9.6}$$

uniformly in $\boldsymbol{\beta} \in \Theta_n$ and $\mathbf{X}_1 \in \mathscr{D}$, where the last inequality follows from (9.2) and results in Corollary 9.2.1, for any large enough C and n. (9.6) could be proved in exactly the same manner. (9.4) thus holds where the equality hold if and only if $\mathbf{w} = \mathbf{0}$, $\mathbf{U} = \mathbf{0}$, and $\mathbf{V} = \mathbf{0}$.

We now move on to the penalty term. First note that we needn't be concerned with terms indexed by $k \notin S_1 \cup S_2$, for in these cases $\mathbf{f}^0_{(k)} = \mathbf{0}$, and $a^0_k = 0$ and the penalty functions are all positive except at the origin. For $k \in S_2$, we have $\mathbf{f}^0_{(k)} = \mathbf{0}$, and

$$
\begin{aligned}
&|a^0_k| + |\mathbf{f}^0_{(k)}| = |a^0_k| = m_k > 0, \\
&|a^0_k + C\alpha_n w_k| + |\mathbf{f}^0_{(k)} + C\alpha_n \mathbf{u}_{(k)}| > \frac{|a^0_k|}{2} = \frac{m_k}{2} > 0,
\end{aligned}
$$

whence as $\max\{\lambda_k : k \in S_1 \cup S_2\} \to 0$,

$$
p_{\lambda_k}(|a^0_k| + |\mathbf{f}^0_{(k)}|) - p_{\lambda_k}(|a^0_k + C\alpha_n w_k| + |\mathbf{f}^0_{(k)} + C\alpha_n \mathbf{u}_{(k)}|) = 0. \quad (9.7)
$$

For $k \in S_1$, by SLLN, we have

$$
\begin{aligned}
&|a^0_k| + |\mathbf{f}^0_{(k)}| \geq \frac{1}{n-1} \sum_{j=2}^{n} |f_k(\mathbf{X}_j)| \to m_k > 0 \ a.s., \\
&|a^0_k + C\alpha_n w_k| + |\mathbf{f}^0_{(k)} + C\alpha_n \mathbf{u}_{(k)}| > \frac{|a^0_k| + |\mathbf{f}^0_{(k)}|}{2} > \frac{m_k}{2} \ a.s., \\
&|\mathbf{f}^0_{(k)}| > 0, \quad |\mathbf{f}^0_{(k)} + C\alpha_n \mathbf{u}_{(k)}| > 0.
\end{aligned}
$$

Therefore, we have

$$p_{\lambda_k}(|a_k^0| + |\mathbf{f}_{(k)}^0|) - p_{\lambda_k}(|a_k^0 + C\alpha_n w_k| + |\mathbf{f}_{(k)}^0 + C\alpha_n \mathbf{u}_{(k)}|) = 0, \tag{9.8}$$

$$p_{\lambda_{k+d}}(|\mathbf{f}_{(k)}^0 + C\alpha_n \mathbf{u}_{(k)}|) - p_{\lambda_{k+d}}(|\mathbf{f}_{(k)}^0 + C\alpha_n \mathbf{u}_{(k)}|) = 0, \tag{9.9}$$

(9.3) thus follows from (9.4) and (9.7)-(9.9). ■

Lemma 9.1.2. *Under conditions in Theorem 6.2.1, we have, with probability one for large enough n*

$$|\hat{\mathbf{f}}_{(k)}(\tilde{\boldsymbol{\beta}})| = 0, \text{ for any } k \notin S_1; \quad \hat{a}_k(\tilde{\boldsymbol{\beta}}) = 0, \text{ for any } k \notin S_1 \cup S_2.$$

Proof of Lemma 9.1.2 Let $\mathbf{a} = (a_k, k = 0, \cdots, d-1)$, $\mathbf{F} = (f_{jk},\ j = 2, \cdots, n, k = 0, \cdots, d-1) = (\mathbf{f}_1^\top, \cdots, \mathbf{f}_n^\top)^\top = (\mathbf{f}_{(0)}, \cdots, \mathbf{f}_{(d-1)})$, and $\mathbf{B} = (b_{jk},\ j = 1, \cdots, n, k = 0, \cdots, d-1) = (\mathbf{b}_1^\top, \cdots, \mathbf{b}_n^\top)^\top = (\mathbf{b}_{(0)}, \cdots, \mathbf{b}_{(d-1)})$ stand for any generic $d \times (1)$ vector, $(n-1) \times d$ and $n \times d$ matrices such that

$$|\mathbf{B} - \mathbf{B}^0|_\infty = \mathcal{O}(\delta_n + h_n^3 + |\delta_{\tilde{\boldsymbol{\beta}}}|^2), \quad |\mathbf{a} - \mathbf{a}^0|_\infty = \mathcal{O}(\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}|),$$
$$|\mathbf{F} - \mathbf{F}^0|_\infty = \mathcal{O}(\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}|). \tag{9.10}$$

In view of Lemma 9.1.1, it suffices to show that with probability one, there exists some small $\epsilon_n > 0$, such that for any such

$\mathbf{a}, \mathbf{F}, \mathbf{B}$,

$$\begin{aligned} \text{for } k \notin S_1, \quad & \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial f_{jk}} < 0, \ \text{ if } f_{jk} \in (-\epsilon_n, 0), \\ & \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial f_{jk}} > 0, \ \text{ if } f_{jk} \in (0, \epsilon_n); \end{aligned} \tag{9.11}$$

$$\begin{aligned} \text{for } k \notin S_1 \cup S_2, \quad & \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial a_k} < 0, \ \text{ if } a_k \in (-\epsilon_n, 0), \\ & \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial a_k} > 0, \ \text{ if } a_k \in (0, \epsilon_n). \end{aligned} \tag{9.12}$$

To prove (9.11), first note that

$$\begin{aligned} \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial f_{jk}} \ = \ & \frac{1}{n^2} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top (\mathbf{a} + \mathbf{f}_j) - (\mathbf{X}_{ij}^\top \boldsymbol{\beta}/h_n) \tilde{\mathbf{X}}_i^\top \mathbf{b}_j\} \\ & \times X_{ik} K_{h_n}(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) + \dot{p}_{\lambda_{d+k}}(|\mathbf{f}_{(k)}|) \text{sign}\,(f_{jk}) \\ & + \dot{p}_{\lambda_k}(|a_k| + |\mathbf{f}_{(k)}|) \text{sign}\,(f_{jk}), \end{aligned} \tag{9.13}$$

where for the first term on the RHS of (9.13), we have that

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \{Y_i - \tilde{\mathbf{X}}_i^\top (\mathbf{a} + \mathbf{f}_j) - (\mathbf{X}_{ij}^\top \boldsymbol{\beta}/h_n) \tilde{\mathbf{X}}_i^\top \mathbf{b}_j\} X_{ik} K_{h_n}(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) \\ & \qquad = \mathscr{O}\Big(\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}| \Big| \mathbf{X}_j \in \mathscr{D}, \boldsymbol{\beta} \in \boldsymbol{\Theta}_n\Big). \end{aligned} \tag{9.14}$$

as a result of (9.2), Corollary 9.2.1 and (9.10). As for the penalty terms in (9.13), it holds that

$$\begin{aligned} & k \in S_2 : \ |a_k| + |\mathbf{f}_{(k)}| = |a_k^0 + C\alpha_n| + C\alpha_n > 0 \to \dot{p}_{\lambda_k}(|a_k| + |\mathbf{f}_{(k)}|) = 0; \\ & k \notin S_1 \cup S_2 : \ |a_k| + |\mathbf{f}_{(k)}| = O(\alpha_n) = o(\lambda_{k+d}) \to \dot{p}_{\lambda_k}(|a_k| + |\mathbf{f}_{(k)}|) = \lambda_k. \end{aligned}$$

Meanwhile $\mathbf{f}_{(k)} = O(\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}|) = o(\lambda_{k+d})$, under condition (6.5), whence $\dot{p}_{\lambda_{k+d}}(|\mathbf{f}_{(k)}|) = \lambda_{k+d}$. This together with (9.14) and (9.13) yields

$$k \in S_2: \ \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial f_{jk}} = \lambda_{k+d}\{\text{sign}\,(f_{jk}) + o(1)\},$$

$$k \notin S_1 \cup S_2: \ \frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial f_{jk}} = (\lambda_k + \lambda_{k+d})\{\text{sign}\,(f_{jk}) + o(1)\},$$

where the term $o(1)$ is uniform in $\mathbf{X}_j \in \mathscr{D}$, $\boldsymbol{\beta} \in \boldsymbol{\Theta}_n$. This finishes the proof of (9.11).

To prove (9.12), we only need to note that for $k \notin S_1 \cup S_2$, $|a_k^0| + |\mathbf{f}_{(k)}^0| = 0$, hence $|a_k| + |\mathbf{f}_{(k)}| = \mathscr{O}(\tau_n + |\delta_{\tilde{\boldsymbol{\beta}}}) = o(\lambda_k)$ and $\dot{p}_{\lambda_k}(|a_k| + |\mathbf{f}_{(k)}|) = \lambda_k$. Consequently,

$$\begin{aligned}
\frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial a_k} &= -\frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a}+\mathbf{f}_j) - (\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_j\} \\
&\quad \times \mathbf{X}_{ik}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) \\
&\quad -\frac{1}{n^2}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top\mathbf{a} - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1\} \\
&\quad \times \mathbf{X}_{ik}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta}) \\
&\quad + \dot{p}_{\lambda_k}(|a_k| + |\mathbf{f}_{(k)}|)\text{sign}\,(a_k) \\
&= \mathscr{O}(\alpha_n|\boldsymbol{\beta} \in \Theta_n) + \lambda_k\text{sign}\,(a_k) \\
&= \lambda_k\{\text{sign}\,(a_k) + o(1)\}.
\end{aligned}$$

The proof is thus complete. ■

For any $\mathbf{x} \in \mathscr{D}$ and $\boldsymbol{\beta} \in \Theta_n$, define

$$
\begin{aligned}
M_3(\mathbf{x}|\boldsymbol{\beta}) &= E[\tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_0(\mathbf{x}) \tilde{X}_i \mathbf{X}_{ix}^\top | \mathbf{X}_i^\top \boldsymbol{\beta}_0 = \mathbf{x}^\top \boldsymbol{\beta}], \\
M_{3(1)}(\mathbf{x}|\boldsymbol{\beta}) &= E[\tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_0(\mathbf{x}) \tilde{X}_{i(1)} \mathbf{X}_{ix}^\top | \mathbf{X}_i^\top \boldsymbol{\beta}_0 = \mathbf{x}^\top \boldsymbol{\beta}], \\
M_{3(2)}(\mathbf{x}|\boldsymbol{\beta}) &= E[\tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_0(\mathbf{x}) \tilde{X}_{i(2)} \mathbf{X}_{ix}^\top | \mathbf{X}_i^\top \boldsymbol{\beta}_0 = \mathbf{x}^\top \boldsymbol{\beta}].
\end{aligned}
$$

Now we give the asymptotics regarding $\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}})$, $\hat{\mathbf{F}}(\tilde{\boldsymbol{\beta}})$, and $\hat{\mathbf{B}}(\tilde{\boldsymbol{\beta}})$, the minimizer of (6.2). Seeing Lemma 9.1.2, let $\hat{\mathbf{a}}_{(2)}(\tilde{\boldsymbol{\beta}})$ denote the subvector of $\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}})$ indexed by S_2, and $\hat{\mathbf{f}}_{j(1)}(\tilde{\boldsymbol{\beta}})$, the subvector of $\hat{\mathbf{f}}_j(\tilde{\boldsymbol{\beta}})$ indexed by S_1.

Lemma 9.1.3. *Under conditions in Theorem 6.2.1, we have for* $j = 1, \cdots, n$,

$$
\begin{aligned}
\hat{\mathbf{a}}_{(2)}(\tilde{\boldsymbol{\beta}}) - \mathbf{a}_{(2)}^0 &= M_0^{-1} \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta}) \, [\tilde{X}_{i(2)} - (\Omega_{21}\Omega_{11}^{-1})(\mathbf{X}_i|\boldsymbol{\beta}) \tilde{X}_{i(1)}] \\
&\quad + o_p(n^{-1/2} + \tau_n |\delta_{\boldsymbol{\beta}}|), \qquad (9.15) \\
\hat{\mathbf{f}}_{j(1)}(\tilde{\boldsymbol{\beta}}) - \mathbf{f}_{j(1)}^0 &= [f.\Omega_{11}]^{-1}(\mathbf{X}_j|\boldsymbol{\beta}_0) \frac{1}{n} \sum_{i=1}^n \varepsilon_i \tilde{X}_{i(1)} K_{h_n}(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) \\
&\quad + \Omega_{11}^{-1}(\mathbf{X}_j|\boldsymbol{\beta}) M_{3(1)}(\mathbf{X}_j|\boldsymbol{\beta}) \delta_{\boldsymbol{\beta}} + \frac{1}{2} h_n^2 \ddot{\mathbf{f}}_0(\mathbf{X}_j) \\
&\quad + \mathscr{O}(h_n \tau_n + |\delta_{\boldsymbol{\beta}}| h_n | \mathscr{D}, \Theta_n), \qquad (9.16) \\
\hat{\mathbf{b}}_j(\tilde{\boldsymbol{\beta}}) - \mathbf{b}_j^0 &= [\Omega(\mathbf{X}_j|\boldsymbol{\beta})]^{-1} \frac{1}{n} \sum_{i=1}^n K_{h_n}(\mathbf{X}_{ij}^\top \boldsymbol{\beta}) (\mathbf{X}_{ij}^\top \boldsymbol{\beta}/h_n) \tilde{\mathbf{X}}_i \varepsilon_i \\
&\quad + \mathscr{O}(h_n \tau_n + h_n |\delta_{\boldsymbol{\beta}}|). \qquad (9.17)
\end{aligned}
$$

Proof of Lemma 9.1.3 For ease of composition, $(\tilde{\boldsymbol{\beta}})$ is left out in $\hat{\mathbf{a}}_{(2)}(\tilde{\boldsymbol{\beta}})$, $\hat{\mathbf{f}}_{j(1)}(\tilde{\boldsymbol{\beta}})$ and $\hat{\mathbf{b}}_j(\tilde{\boldsymbol{\beta}})$, so that these are replaced with

$\hat{\mathbf{a}}_{(2)}$, $\hat{\mathbf{f}}_{j(1)}$ and $\hat{\mathbf{b}}_j$, respectively. Nevertheless we should remember these estimates depend on $\tilde{\boldsymbol{\beta}}$. Based on the results in Lemma 9.1.2, we have with probability one, that among the elements of $\hat{\mathbf{a}} = (\hat{a}_0, \cdots, \hat{a}_{d-1})^\top$, $\hat{a}_k = 0$, if $k \notin S_1 \cup S_2$; while for $\hat{\mathbf{f}}_j = (\hat{f}_{j0}, \cdots, \hat{f}_{jd-1})^\top$, $\hat{f}_{jk} = 0$ if $k \notin S_1$. Now with $\Delta_{ij} = Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a}^0 + \mathbf{f}_j^0) - (\mathbf{X}_{ij}^\top \boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top \mathbf{b}_j^0$, we have any $k \in S_1 \cup S_2$,

$$
\begin{aligned}
\frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial a_k} \;=\; & -\frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{\Delta_{ij} - \tilde{\mathbf{X}}_i^\top(\mathbf{a} - \mathbf{a}^0 + \mathbf{f}_j - \mathbf{f}_j^0) \\
& -(\mathbf{X}_{ij}^\top \boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top(\mathbf{b}_j - \mathbf{b}_j^0)\}\mathbf{X}_{ik}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) \\
& +\frac{1}{n^2}\sum_{i=1}^{n}\{\Delta_{i1} - \tilde{\mathbf{X}}_i^\top(\mathbf{a} - \mathbf{a}^0) - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n) \\
& \times\tilde{\mathbf{X}}_i^\top(\mathbf{b}_j - \mathbf{b}_1^0)\}\mathbf{X}_{ik}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta}) \\
& +\dot{p}_{\lambda_k}(|a_k| + |\mathbf{f}_{(k)}|)\mathrm{sign}\,(a_k).
\end{aligned}
\tag{9.18}
$$

Regarding the penalty term within (9.18), since $|a_k^0| + |\mathbf{f}_{(k)}^0| > 0$ for $k \in S_1 \cup S_2$, then according to Lemma 9.1.1, it is also true that $|\hat{a}_k| + |\hat{\mathbf{f}}_{(k)}| > 0$; whence $\dot{p}_{\lambda_{1k}}(|\hat{a}_k| + |\hat{\mathbf{f}}_{(k)}|) = 0$, since $\max\{\lambda_k, k \in S_1 \cup S_2\} = o(1)$. Therefore, $\hat{\mathbf{a}}(\boldsymbol{\beta}), \hat{\mathbf{F}}(\boldsymbol{\beta}), \hat{\mathbf{B}}(\boldsymbol{\beta})$ must satisfy the following equation

$$
\begin{aligned}
& \frac{1}{n^2}\sum_{j=2}^{n}\sum_{i=1}^{n}\{\Delta_{ij} - \tilde{\mathbf{X}}_i^\top(\hat{\mathbf{a}} - \mathbf{a}^0 + \hat{\mathbf{f}}_j - \mathbf{f}_j^0) - (\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top(\hat{\mathbf{b}}_j - \mathbf{b}_j^0)\} \\
& \times\mathbf{X}_{ik}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) + \frac{1}{n^2}\sum_{i=1}^{n}\{\Delta_{ij} - \tilde{\mathbf{X}}_i^\top(\hat{\mathbf{a}} - \mathbf{a}^0) - (\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top(\hat{\mathbf{b}}_1 - \mathbf{b}_1^0)\} \\
& \times\mathbf{X}_{ik}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta}) = 0, \qquad k \in S_1 \cup S_2,
\end{aligned}
$$

the matrix form of which is such that

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=1}^{n}\Delta_{ij}\tilde{\mathbf{X}}_{i(2)}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}) &= \sum_{j=1}^{n}\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})\tilde{\mathbf{X}}_{i(2)}\tilde{\mathbf{X}}_{i(2)}^{\top}(\hat{\mathbf{a}}_{(2)}-\mathbf{a}_{(2)}^{0})\\ &+\sum_{j=2}^{n}[\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})\mathbf{X}_{i(2)}\tilde{\mathbf{X}}_{i(1)}^{\top}](\hat{\mathbf{f}}_{j(1)}-\mathbf{f}_{j(1)}^{0})\\ &+\sum_{j=1}^{n}[\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})(\mathbf{X}_{i1}^{\top}\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_{i(2)}\tilde{\mathbf{X}}_{i}^{\top}]\\ &\times(\hat{\mathbf{b}}_j-\mathbf{b}_j^{0}). \qquad (9.19)\end{aligned}$$

We now move on to the $\hat{\mathbf{f}}_j$s. Note that for $k\in S_1$, $|\mathbf{f}_{(k)}^{0}|>0$, whence $|\hat{\mathbf{f}}_{(k)}|=|\mathbf{f}_{(k)}^{0}+O(\alpha_n)|>0$ and $|\hat{\alpha}_k|+|\hat{\mathbf{f}}_{(k)}|>0$. Therefore, $\hat{\mathbf{a}}(\boldsymbol{\beta}),\hat{\mathbf{F}}(\boldsymbol{\beta}),\hat{\mathbf{B}}(\boldsymbol{\beta})$ must also satisfy the following equation

$$\begin{aligned}\frac{\partial Q_n(\mathbf{a},\mathbf{F},\mathbf{B}|\boldsymbol{\beta},\boldsymbol{\lambda})}{\partial f_{jk}} &= \frac{1}{n^2}\sum_{i=1}^{n}\{\Delta_{ij}-\tilde{\mathbf{X}}_i^{\top}(\mathbf{a}-\mathbf{a}^{0}+\mathbf{f}_j-\mathbf{f}_j^{0})\\ &-(\mathbf{X}_{i1}^{\top}\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^{\top}(\mathbf{b}_j-\mathbf{b}_j^{0})\}X_{ik}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})=0,\end{aligned}$$

for $k\in S_1$, the matrix form of which is such that for $j=2,\cdots,n$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Delta_{ij}\tilde{X}_{i(1)}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}) &= \Big\{\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})\tilde{X}_{i(1)}\tilde{\mathbf{X}}_{i(2)}^{\top}\Big\}(\hat{\mathbf{a}}_{(2)}-\mathbf{a}_{(2)}^{0})\\ &+\Big\{\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})\tilde{X}_{i(1)}\tilde{\mathbf{X}}_{i(1)}^{\top}\Big\}(\hat{\mathbf{f}}_{j(1)}-\mathbf{f}_{j(1)}^{0})\\ &+[\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})\Big(\frac{\mathbf{X}_{i1}^{\top}\boldsymbol{\beta}}{h_n}\Big)\tilde{X}_{i(1)}\tilde{\mathbf{X}}_{i}^{\top}]\\ &\times(\hat{\mathbf{b}}_j-\mathbf{b}_j^{0}). \qquad (9.20)\end{aligned}$$

Since $\hat{\mathbf{b}}_j - \mathbf{b}_j^0 = \mathcal{O}(\tau_n + |\delta_{\boldsymbol{\beta}}|)$ (uniformly in j), from (9.20), we have, for $j = 2, \cdots, n$,

$$\begin{aligned}\hat{\mathbf{f}}_{j(1)} - \mathbf{f}_{j(1)}^0 &= [S_{11,j}]^{-1}\{T_{1j} - S_{12,j}(\hat{\mathbf{a}}_{(2)} - \mathbf{a}_{(2)}^0)\} \\ &\quad + \mathcal{O}(h_n\tau_n + h_n|\delta_{\boldsymbol{\beta}}|\Big|\mathscr{D}, \Theta_n), \qquad (9.21)\end{aligned}$$

where

$$\begin{aligned}S_{11,j} &= \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta})\tilde{X}_{i(1)}\tilde{\mathbf{X}}_{i(1)}^\top \\ &= \Omega_{11}(\mathbf{X}_j|\boldsymbol{\beta})f(\mathbf{X}_j|\boldsymbol{\beta}) + (O)(\tau_n|\mathscr{D}, \Theta_n) \\ S_{22,j} &= \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta})\tilde{X}_{i(2)}\tilde{\mathbf{X}}_{i(2)}^\top \\ &= \Omega_{22}(\mathbf{X}_j|\boldsymbol{\beta})f(\mathbf{X}_j|\boldsymbol{\beta}) + (O)(\tau_n|\mathscr{D}, \Theta_n) \\ S_{12,j} &= \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta})\tilde{X}_{i(1)}\tilde{\mathbf{X}}_{i(2)}^\top \\ &= \Omega_{12}(\mathbf{X}_j|\boldsymbol{\beta})f(\mathbf{X}_j|\boldsymbol{\beta}) + (O)(\tau_n|\mathscr{D}, \Theta_n) \\ T_{1j} &= \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij}\tilde{X}_{i(1)}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}), \\ T_{2j} &= \frac{1}{n}\sum_{i=1}^{n} \Delta_{ij}\tilde{X}_{i(2)}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}).\end{aligned}$$

Plug (9.21) into (9.19), we have with $S_{21,j} = S_{12,j}^\top$,

$$\begin{aligned}\hat{\mathbf{a}}_{(2)} - \mathbf{a}_{(2)}^0 &= \left\{\frac{1}{n}\sum_{j=1}^{n} S_{22,j}\right\}^{-1}\frac{1}{n}\sum_{j=1}^{n}(T_{2j} - S_{21,j}[S_{11,j}]^{-1}T_{1j}) \\ &= \left\{\frac{1}{n}\sum_{j=1}^{n} S_{22,j}\right\}^{-1}\frac{1}{n}\sum_{j=1}^{n}\Delta_{ij}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta}) \\ &\quad \times(\tilde{X}_{i(2)} - S_{21,j}[S_{11,j}]^{-1}\tilde{X}_{i(1)}). \qquad (9.22)\end{aligned}$$

To quantify the term on the RHS of (9.22), we make use of the following facts

$$\begin{aligned}\Delta_{ij} &= Y_i - \tilde{\mathbf{X}}_i^\top(\mathbf{a}^0 + \mathbf{f}_j^0) - (\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_j^0 \\ &= \varepsilon_i + \tilde{\mathbf{X}}_i^\top\dot{\mathbf{f}}_{nj}\mathbf{X}_{ij}^\top\delta_{\boldsymbol{\beta}}/h_n + \frac{1}{2}h_n^2\tilde{\mathbf{X}}_i^\top\ddot{\mathbf{f}}_0(\mathbf{X}_j)(\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)^2 \\ &\quad + \frac{1}{2}\tilde{\mathbf{X}}_i^\top\ddot{\mathbf{f}}_0(\mathbf{X}_j)\delta_{\boldsymbol{\beta}}^\top\mathbf{X}_{ij}\mathbf{X}_{ij}^\top\delta_{\boldsymbol{\beta}} + h_n\tilde{\mathbf{X}}_i^\top\ddot{\mathbf{f}}_0(\mathbf{X}_j)(\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\mathbf{X}_{ij}^\top\delta_{\boldsymbol{\beta}} \\ &\quad + \mathscr{O}(|\delta_{\boldsymbol{\beta}}|^3 + h_n^3|\mathscr{D},\ \boldsymbol{\Theta}_n);\end{aligned}$$

$$\begin{aligned}&\frac{1}{n^2}\sum_{j,i=1}^n \varepsilon_i K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta})\{\tilde{X}_{i(2)} - S_{21,j}[S_{11,j}]^{-1}\tilde{X}_{i(1)}\} \\ &= \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta})\Big[\tilde{X}_{i(2)} - \Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta})\{\Omega_{11}(\mathbf{X}_i|\boldsymbol{\beta})\}^{-1}\tilde{X}_{i(1)}\Big] \\ &+ o_p(n^{-1/2});\end{aligned} \tag{9.23}$$

and

$$\begin{aligned}&\frac{1}{n^2}\sum_{j,i=1}^n K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta})\{\tilde{X}_{i(2)} - S_{21,j}[S_{11,j}]^{-1}\tilde{X}_{i(1)}\}\tilde{\mathbf{X}}_i^\top\dot{\mathbf{f}}_0(\mathbf{X}_j)\mathbf{X}_{ij}^\top\delta_{\boldsymbol{\beta}} \\ &= \frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i|\boldsymbol{\beta})[\tilde{X}_{i(2)} - \Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta})\{\Omega_{11}(\mathbf{X}_i|\boldsymbol{\beta})\}^{-1}\tilde{X}_{i(1)}]\tilde{\mathbf{X}}_i^\top\dot{\mathbf{f}}_0(\mathbf{X}_i) \\ &\times\nu^\top(\mathbf{X}_i|\boldsymbol{\beta})\delta_{\boldsymbol{\beta}} + O_p(\tau_n|\delta_{\boldsymbol{\beta}}|) \\ &= O_p(\tau_n|\delta_{\boldsymbol{\beta}}|);\end{aligned} \tag{9.24}$$

where the last equality holds due to the fact that

$$E\Big[f(\mathbf{X}_i|\boldsymbol{\beta})[\tilde{X}_{i(2)}-\Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta})\Omega_{11}^{-1}(\mathbf{X}_i|\boldsymbol{\beta})\tilde{X}_{i(1)}]\tilde{\mathbf{X}}_i^\top\dot{\mathbf{f}}_0(\mathbf{X}_i)\nu^\top(\mathbf{X}_i|\boldsymbol{\beta})\Big]$$
$$=E\Big[f(\mathbf{X}_i|\boldsymbol{\beta})E\{[\tilde{X}_{i(2)}-\Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta})\Omega_{11}^{-1}(\mathbf{X}_i|\boldsymbol{\beta})\tilde{X}_{i(1)}]\tilde{\mathbf{X}}_{i(1)}^\top|\mathbf{X}_i^\top\boldsymbol{\beta}\}\dot{\mathbf{f}}_0(\mathbf{X}_i)$$
$$\times\nu^\top(\mathbf{X}_i|\boldsymbol{\beta})\Big]=\mathbf{0}.$$

Similarly,

$$\frac{1}{n^2}\sum_{j,i=1}^{n}K_{h_n}(\mathbf{X}_{ij}^\top\boldsymbol{\beta})\{\tilde{X}_{i(2)}-S_{21,j}[S_{11,j}]^{-1}\tilde{X}_{i(1)}\}\tilde{\mathbf{X}}_i^\top\ddot{\mathbf{f}}_0(\mathbf{X}_j)(\mathbf{X}_{i\mathbf{x}}^\top\boldsymbol{\beta}/h_n)^2$$
$$=\frac{1}{n}\sum_{i=1}^{n}f(\mathbf{X}_i|\boldsymbol{\beta})[\tilde{X}_{i(2)}-\Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta})\{\Omega_{11}(\mathbf{X}_i|\boldsymbol{\beta})\}^{-1}\tilde{X}_{i(1)}]\tilde{\mathbf{X}}_i^\top\ddot{\mathbf{f}}_0(\mathbf{X}_i)$$
$$+o_p(\tau_n)=o_p(\tau_n), \tag{9.25}$$

since

$$E[f(\mathbf{X}_i|\boldsymbol{\beta})[\tilde{X}_{i(2)}-\Omega_{21}(\mathbf{X}_i|\boldsymbol{\beta})\{\Omega_{11}(\mathbf{X}_i|\boldsymbol{\beta})\}^{-1}\tilde{X}_{i(1)}]\tilde{\mathbf{X}}_i^\top\ddot{\mathbf{f}}_0(\mathbf{X}_i)]=\mathbf{0}.$$

From (9.23)-(9.25), we have (9.15), which means $\hat{\mathbf{a}}_{(2)}-\mathbf{a}_{(2)}^0$ of order $O_p(n^{-1/2})$. This together with (9.22) leads to

$$\hat{\mathbf{f}}_{j(1)}-\mathbf{f}_{j(1)}^0=[S_{11,j}]^{-1}T_{1j}+\mathscr{O}(h_n\tau_n+h_n|\delta_{\boldsymbol{\beta}}|\Big|\mathscr{D},\Theta_n),$$

from which (9.16) easily follows. Lastly,with the derivatives with respect to $\mathbf{B}$ being zero, $\hat{\mathbf{a}}(\boldsymbol{\beta}),\hat{\mathbf{F}}(\boldsymbol{\beta}),\hat{\mathbf{B}}(\boldsymbol{\beta})$ must also satisfy

the following that for $j = 2, \cdots, n$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i\Delta_{ij} \\
&= \Big\{\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_{i(2)}^{\top}\Big\}(\hat{\mathbf{a}}_{(2)} - \mathbf{a}_{(2)}^{0}) \\
&+\Big\{\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_{i(1)}^{\top}\Big\}(\hat{\mathbf{f}}_{j(1)} - \mathbf{f}_{j(1)}^{0}) \\
&+\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta})(\mathbf{X}_{ij}^{\top}\boldsymbol{\beta}/h_n)^2\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^{\top}\Big\}](\hat{\mathbf{b}}_j - \mathbf{b}_j^{0}), \qquad (9.26)
\end{aligned}$$

and also

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{i1}^{\top}\boldsymbol{\beta})\Big(\frac{\mathbf{X}_{i1}^{\top}\boldsymbol{\beta}}{h_n}\Big)\tilde{\mathbf{X}}_i\Delta_{ij} \\
&= [\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{i1}^{\top}\boldsymbol{\beta})\Big(\frac{\mathbf{X}_{i1}^{\top}\boldsymbol{\beta}}{h_n}\Big)\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_{i(2)}^{\top}](\hat{\mathbf{a}}_{(2)} - \mathbf{a}_{(2)}^{0}) \\
&+\frac{1}{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{i1}^{\top}\boldsymbol{\beta})\Big(\frac{\mathbf{X}_{i1}^{\top}\boldsymbol{\beta}}{h_n}\Big)^2\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^{\top}](\hat{\mathbf{b}}_1 - \mathbf{b}_1^{0}). \qquad (9.27)
\end{aligned}$$

(9.17) is then a result of Lemma 9.1.1, (9.26) and (9.27). The proof is thus complete. ■

We now move on to the study of $\hat{\boldsymbol{\beta}}$, the minimizer of (6.3).

Let $\hat{\mathbf{f}}_1 \equiv \mathbf{0}$, and it is immediately clear that $\hat{\boldsymbol{\beta}}$ minimizes

$$\begin{aligned} Q_n(\boldsymbol{\beta}|\boldsymbol{\lambda},\tilde{\boldsymbol{\beta}}) &\equiv \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\{Y_i - \tilde{\mathbf{X}}_i^\top(\hat{\mathbf{a}}(\tilde{\boldsymbol{\beta}}) + \hat{\mathbf{f}}_j(\tilde{\boldsymbol{\beta}})) - (\mathbf{X}_{ij}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\hat{\mathbf{b}}_j(\tilde{\boldsymbol{\beta}})\}^2 \\ &\quad \times K_{h_n}(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}}) + \sum_{k=1}^{d} p_{\tilde{\lambda}_k}(|\beta_k|) \\ &= (\boldsymbol{\beta}-\boldsymbol{\beta}_0)^\top S_n(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \{R_n(\tilde{\boldsymbol{\beta}})\}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}_0) \\ &\quad + \sum_{k=1}^{d} p_{\tilde{\lambda}_k}(|\boldsymbol{\beta}_k|), \end{aligned} \tag{9.28}$$

where

$$\begin{aligned} S_n(\tilde{\boldsymbol{\beta}}) &= \frac{1}{n^2h_n^2}\sum_{j=1}^{n}\sum_{i=1}^{n} K_{h_n}(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}})(\tilde{\mathbf{X}}_i^\top\hat{\mathbf{b}}_j)^2\mathbf{X}_{ij}\mathbf{X}_{ij}^\top, \\ R_n(\tilde{\boldsymbol{\beta}}) &= \frac{2}{n^2h_n}\sum_{i,j=1}^{n} K_{h_n}(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}})\{Y_i - \tilde{\mathbf{X}}_i^\top(\hat{\mathbf{a}}+\hat{\mathbf{f}}_j) - (\mathbf{X}_{ij}^\top\boldsymbol{\beta}_0/h_n) \\ &\quad \times\tilde{\mathbf{X}}_i^\top\hat{\mathbf{b}}_j\}\tilde{\mathbf{X}}_i^\top\hat{\mathbf{b}}_j\mathbf{X}_{ij}. \end{aligned}$$

Using results in Lemma 9.1.3, we understand that

$$\begin{aligned} S_n(\tilde{\boldsymbol{\beta}}) &= E[f(\mathbf{X}|\boldsymbol{\beta}_0)C(\mathbf{X}|\boldsymbol{\beta}_0)] + \mathcal{O}(|\delta_{\tilde{\boldsymbol{\beta}}}| + \tau_n|\Theta_n) \\ &= C_0 + \mathcal{O}(|\delta_{\tilde{\boldsymbol{\beta}}}| + \tau_n|\Theta_n), \end{aligned} \tag{9.29}$$

$$\begin{aligned} R_n(\tilde{\boldsymbol{\beta}}) &= \frac{2}{n}\sum_i \varepsilon_i\nu(\mathbf{X}_i|\boldsymbol{\beta})\tilde{\mathbf{X}}_i^\top\hat{\mathbf{b}}_i - \frac{2}{n}\sum_i M_{3(1)}^\top(\mathbf{X}_i|\boldsymbol{\beta})\Omega_{11}^{-1}(\mathbf{X}_i|\boldsymbol{\beta})\tilde{\mathbf{X}}_{i(1)} \\ &\quad \times f(\mathbf{X}_i|\boldsymbol{\beta})\varepsilon_i - \frac{2}{n}E\{(f.M_{3(2)}^\top)(\mathbf{X}|\boldsymbol{\beta})\}M_0^{-1}\Big\{\sum_{i=1}^{n}\varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta}) \\ &\quad \times [\tilde{X}_{i(2)} - (\Omega_{21}\Omega_{11}^{-1})(\mathbf{X}_i|\boldsymbol{\beta})\tilde{X}_{i(1)}]\Big\} - \frac{2}{n^2h_n}\sum_{i,j} K_{h_n}(\mathbf{X}_{ij}^\top\tilde{\boldsymbol{\beta}}) \\ &\quad \times\tilde{\mathbf{X}}_i^\top\hat{\mathbf{b}}_j\mathbf{X}_{ij}^\top\tilde{\mathbf{X}}_{i(1)}^\top\Omega_{11}^{-1}(\mathbf{X}_j|\boldsymbol{\beta})M_{3(1)}(\mathbf{X}_j|\boldsymbol{\beta})\delta_{\tilde{\boldsymbol{\beta}}} + o(n^{-1/2}|\Theta_n) \\ &= \mathcal{O}((\log n/n)^{1/2} + |\delta_{\tilde{\boldsymbol{\beta}}}|\Big|\Theta_n). \end{aligned} \tag{9.30}$$

The first result states that $\hat{\boldsymbol{\beta}}$ is consistent, if the initial estimator $\tilde{\boldsymbol{\beta}}$ is.

Lemma 9.1.4. *Suppose conditions in Theorem 6.2.1 hold. then*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \mathcal{O}((\log n/n)^{1/2} + |\delta_{\tilde{\boldsymbol{\beta}}}| \Big| \tilde{\boldsymbol{\beta}}).$$

Proof of Lemma 9.1.4 Let $\alpha_n = (\log n/n)^{1/2} + |\delta_{\tilde{\boldsymbol{\beta}}}|$. It suffices to show that for any large enough $C > 0$, such that for any $\mathbf{b} = (b_1, \cdots, b_d)^\top \in R^d$ such that $\mathbf{b}^\top \boldsymbol{\beta}_0 = 0$ and $|\mathbf{b}| = 1$,

$$Q_n(\boldsymbol{\beta}_0 + Ca_n\mathbf{b}|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}}) > Q_n(\boldsymbol{\beta}_0|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}}),$$

which easily follows from the fact that

$$\begin{aligned} & Q(\boldsymbol{\beta}_0 + Ca_n\mathbf{b}|\tilde{\lambda}, \tilde{\boldsymbol{\beta}}) - Q(\boldsymbol{\beta}_0|\tilde{\lambda}, \tilde{\boldsymbol{\beta}}) \\ & = C^2 a_n^2 \mathbf{b}^\top S_n(\tilde{\boldsymbol{\beta}})\mathbf{b} - 2Ca_n\mathbf{b}^\top R_n(\tilde{\boldsymbol{\beta}}) \\ & \quad + \sum_{k=1}^{d} \{p_{\tilde{\lambda}_k}(|\beta_{0k} + Ca_n b_k|) - p_{\tilde{\lambda}_k}(|\beta_{0k}|)\}, \end{aligned}$$

assumption (C4), (9.29) and (9.30) in exactly the same way as in Lemma 9.1.1. The proof is thus complete. ■

Lemma 9.1.5. *Suppose conditions in Theorem 6.2.1 hold. Then with probability one, $\hat{\boldsymbol{\beta}}_{01} = \mathbf{0}$ for large enough n.*

Proof of Lemma 9.1.5 Let $\alpha_n = (\log n/n)^{1/2} + |\delta_{\tilde{\boldsymbol{\beta}}}|$. For any $k = 1, \cdots, d_0$, consider

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\beta}|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}})}{\partial \beta_k} &= \dot{p}_{\tilde{\lambda}_k}(|\beta_k|)\text{sign}(\beta_k) + 2[S_n(\tilde{\boldsymbol{\beta}})]_{(k)}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\ &\quad + [R_n(\tilde{\boldsymbol{\beta}})]_{(k)}, \end{aligned} \tag{9.31}$$

where $[S_n(\tilde{\boldsymbol{\beta}})]_{(k)}$ stands for the kth row of $S_n(\tilde{\boldsymbol{\beta}})$, and $[R_n(\tilde{\boldsymbol{\beta}})]_{(k)}$, the kth element of $R_n(\tilde{\boldsymbol{\beta}})$. Therefore, due to 9.30, for any $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^\top \in R^d$ such that $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O(\alpha_n)$, the last two terms are of order $\mathcal{O}(\alpha_n|\Theta) = o(\tilde{\lambda}_k)$, whence

$$\frac{\partial Q_n(\boldsymbol{\beta}|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}})}{\partial \beta_k} = \tilde{\lambda}_k \{\dot{p}_{\tilde{\lambda}_k}(|\beta_k|)\text{sign}\,(\beta_k)/\tilde{\lambda}_k + o(1)\}.$$

Since $\beta_k = \beta_{0k} + O(\alpha_n) = o(\tilde{\lambda}_k)$, we have $\dot{p}_{\tilde{\lambda}_1}(|\beta_k|) = \tilde{\lambda}_k$ for large enough n. Therefore,

$$\partial Q_n(\boldsymbol{\beta}|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}})/\partial \beta_k > 0, \text{ if } \beta_k > 0; \quad \partial Q_n(\boldsymbol{\beta}|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}})/\partial \beta_k < 0, \text{ if } \beta_k < 0.$$

The proof is thus complete. ■

9.2 Proofs of the main results

Proof of Theorem 6.2.1 Claims in part (a) are as given in Lemma 9.1.2 and Lemma 9.1.5. Those in part (b) follows directly from Lemma 9.1.3 and the root-n consistency of $\hat{\boldsymbol{\beta}}$. Therefore, we need only concentrate on proving part (b). To this aim, first of all as a follow-up on (9.29) and (9.30), we claim that in the asymptotic expression for $\{S_n(\tilde{\boldsymbol{\beta}})\}^{-1}R_n(\tilde{\boldsymbol{\beta}})$, the term which concerns $\delta_{\tilde{\boldsymbol{\beta}}}$ (representing the effect of the initial estimate $\tilde{\boldsymbol{\beta}}$) diminishes geometrically, due to the fact that the (absolute) eigenvalues of $C_0^+ E[M_3^\top(\mathbf{X}|\boldsymbol{\beta})\Omega_{11}^{-1}(\mathbf{X}|\boldsymbol{\beta}_0)M_3(\mathbf{X}|\boldsymbol{\beta})]$

are all strictly less than one for all $\boldsymbol{\beta} \in Theta_n$. This could be argued as follows: by the Cauchy-Schwartz inequality, for any real vectors $\mathbf{a}, \mathbf{b}$ of conformable lengths, we have

$$\{\mathbf{a}^\top M_3(\mathbf{X}|\boldsymbol{\beta})\mathbf{b}\}^2 \leq \mathbf{a}^\top \Omega_{11}(\mathbf{X}|\boldsymbol{\beta}_0)\mathbf{a}\,\mathbf{b}^\top C_0 \mathbf{b},$$

i.e. the (nonzero) eigenvalues of

$$[\Omega_{11}(\mathbf{X}|\boldsymbol{\beta}_0)]^+ M_3(\mathbf{X}|\boldsymbol{\beta})\mathbf{b}\mathbf{b}^\top [M_3(\mathbf{X}|\boldsymbol{\beta})]^\top,$$

identical to those of

$$\mathbf{b}^\top [M_3(\mathbf{X}|\boldsymbol{\beta})]^\top [\Omega_{11}(\mathbf{X}|\boldsymbol{\beta}_0)]^+ M_3(\mathbf{X}|\boldsymbol{\beta})\mathbf{b},$$

are all less than $\mathbf{b}^\top C_0 \mathbf{b}$; since this holds for any $\mathbf{X}$, taking expectation with $\mathbf{X}$, the same conclusion still holds. A direct consequence of this claim is that from now on when dealing with $R_n(\tilde{\boldsymbol{\beta}})$, we could safely ignore this term which involves $\delta_{\tilde{\boldsymbol{\beta}}}$.

For any $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_d)^\top \in R^d$, consider a partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are of length d_0 and $d - d_0$, respectively. According to Lemma (9.1.5), with n large enough, $\hat{\boldsymbol{\beta}}$ as a local maximizer of (9.28), must take the form $\hat{\boldsymbol{\beta}} = (\mathbf{0}, \hat{\boldsymbol{\beta}}_2^\top)^\top$, and consequently satisfy the following normal equation

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\beta}|\boldsymbol{\lambda}, \tilde{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}_2}\Big|_{\boldsymbol{\beta}=(\mathbf{0}, \hat{\boldsymbol{\beta}}_2^\top)^\top} &= [S_n(\tilde{\boldsymbol{\beta}})]_{(2)}(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_{02}) + [R_n(\tilde{\boldsymbol{\beta}})]_{(2)} + \dot{\mathbf{p}}_{\tilde{\lambda}_{(2)}}(|\boldsymbol{\beta}_2|) \\ &= \mathbf{0}; \end{aligned} \tag{9.32}$$

here $[S_n(\tilde{\boldsymbol{\beta}})]_{(2)}$ is the $(d-d_0)\times(d-d_0)$ lower diagonal submatrix of $S_n(\tilde{\boldsymbol{\beta}})$, $[R_n(\tilde{\boldsymbol{\beta}})]_{(2)}$ is the vector consisting the last $d-d_0$ elements of $R_n(\tilde{\boldsymbol{\beta}})$, and $\dot{\mathbf{p}}_{\tilde{\lambda}_{(2)}}(|\boldsymbol{\beta}_2|)$ is a $(d-d_0)\times 1$ vector with elements $\dot{p}_{\tilde{\lambda}_k}(|\beta_k|)\text{sign}(\beta_k)$, $d_0+1\le k\le d$. As $\hat{\boldsymbol{\beta}}_2$ is a consistent estimator of $\boldsymbol{\beta}_{02}$ (Lemma 9.1.4), we have $\dot{\mathbf{p}}_{\tilde{\lambda}_{(2)}}(|\hat{\boldsymbol{\beta}}_2|)=\mathbf{0}$ which together with (9.32) implies that

$$\hat{\boldsymbol{\beta}}_2-\boldsymbol{\beta}_{02}=[S_n(\tilde{\boldsymbol{\beta}})]^{+}_{(2)}[R_n(\tilde{\boldsymbol{\beta}})]_{(2)}.$$

First of all, since $[S_n(\tilde{\boldsymbol{\beta}})]_{(2)}=B_{02}B_{02}^{\top}C_{02}B_{02}B_{02}^{\top}+O(\tau_n+|\delta_{\tilde{\boldsymbol{\beta}}}|)$, where B_{02} is the $(d-d_0)\times(d-d_0-1)$ matrix with orthonormal columns given by the $(d-d_0-1)$ eigen-vectors corresponding to the nonzero eigen-values of C_{02}, we have

$$\begin{aligned}[S_n(\tilde{\boldsymbol{\beta}})_{(2)}]^{+} &= B_{02}(B_{02}^{\top}C_{02}B_{02})^{+}B_{02}^{\top}+O(\tau_n+|\delta_{\tilde{\boldsymbol{\beta}}}|)\\ &= C_{02}^{+}+O(\tau_n+|\delta_{\tilde{\boldsymbol{\beta}}}|),\end{aligned}$$

and this together with (9.30) leads to

$$\begin{aligned}[S_n(\tilde{\boldsymbol{\beta}})]^{+}[R_n(\tilde{\boldsymbol{\beta}})]_{(2)} &= \frac{2}{n}C_{02}^{+}\sum_i \varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta})\nu_{(2)}(\mathbf{X}_i|\boldsymbol{\beta})\tilde{\mathbf{X}}_i^{\top}\hat{\mathbf{b}}_i\\ &\quad+\mathcal{O}\left(\delta_n^2/h_n+h_n|\delta_{\tilde{\boldsymbol{\beta}}}|\middle|\tilde{\boldsymbol{\beta}}\right)\\ &\quad-\frac{2}{n}C_{02}^{+}\sum_i M_{(1)(2)}(\mathbf{X}_i|\boldsymbol{\beta})\Omega_{11}^{-1}(\mathbf{X}_i|\boldsymbol{\beta})\tilde{\mathbf{X}}_{i(1)}f(\mathbf{X}_i|\boldsymbol{\beta})\varepsilon_i\\ &\quad-\frac{2}{n}C_{02}^{+}E\{(f.M_{(2)(2)}^{\top})(\mathbf{X}|\boldsymbol{\beta})\}M_0^{-1}\\ &\quad\times\left\{\sum_{i=1}^{n}\varepsilon_i f(\mathbf{X}_i|\boldsymbol{\beta})\,[\tilde{X}_{i(2)}-(\Omega_{21}\Omega_{11}^{-1})(\mathbf{X}_i|\boldsymbol{\beta})\tilde{X}_{i(1)}]\right\},\end{aligned}$$

where we have left out the term which involves $\delta_{\tilde{\boldsymbol{\beta}}}$ in (9.30), as argued at the beginning of the proof. Part (b) thus follows easily from (9.33), and the continuity in $\boldsymbol{\beta}$ of functions such as $f(\mathbf{X}_i|\boldsymbol{\beta})$. ■

Corollary 9.2.1. *Under conditions (C1)-(C4), we have*

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\tilde{\mathbf{X}}_i^\top\mathbf{f}_0(\mathbf{x})-\tilde{\mathbf{X}}_i^\top\dot{\mathbf{f}}_{n0}(\mathbf{x})(\mathbf{X}_{i\mathbf{x}}^\top\boldsymbol{\beta}/h_n)\}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})$$
$$=\frac{1}{n}\sum_{i=1}^{n}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\varepsilon_i+\mathscr{O}(\tau_n|\delta_{\boldsymbol{\beta}}||\mathscr{D},\ \Theta_n),$$
$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\tilde{\mathbf{X}}_i^\top(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\{Y_i-\tilde{\mathbf{X}}_i^\top\mathbf{a}^0-(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1^0\}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})$$
$$=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\tilde{\mathbf{X}}_i^\top((\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\varepsilon_i+\mathscr{O}(\tau_n|\delta_{\boldsymbol{\beta}}||\mathscr{D},\ \Theta_n)=\mathscr{O}(\tau_n|\mathscr{D},\ \Theta_n),$$
$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\{Y_i-\tilde{\mathbf{X}}_i^\top\mathbf{a}^0-(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1^0\}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})$$
$$=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\varepsilon_i+\mathscr{O}(\tau_n|\delta_{\boldsymbol{\beta}}||\mathscr{D},\ \Theta_n)=\mathscr{O}(\tau_n|\mathscr{D},\ \Theta_n),$$
$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathbf{X}}_i\{Y_i-\tilde{\mathbf{X}}_i^\top\mathbf{a}^0-(\mathbf{X}_{i1}^\top\boldsymbol{\beta}/h_n)\tilde{\mathbf{X}}_i^\top\mathbf{b}_1^0\}K_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})$$
$$=\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iK_{h_n}(\mathbf{X}_{i1}^\top\boldsymbol{\beta})\varepsilon_i+\mathscr{O}(|\delta_{\boldsymbol{\beta}}|+h_n^2|\mathscr{D},\ \Theta_n)=\mathscr{O}(|\delta_{\boldsymbol{\beta}}|+\tau_n|\mathscr{D},\ \Theta_n).$$

Proof of Corollary 9.2.1 These are standard results in ker-

nel smoothing, which follow easily from the fact that

$$
\begin{aligned}
& Y_i - \tilde{\mathbf{X}}_i^\top \mathbf{f}_0(\mathbf{x}) - \tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_{n0}(\mathbf{x})(\mathbf{X}_{i\mathbf{x}}^\top \boldsymbol{\beta}/h_n) \\
& = \varepsilon_i + \tilde{\mathbf{X}}_i^\top \dot{\mathbf{f}}_{n0}(\mathbf{x}) \mathbf{X}_{i\mathbf{x}}^\top \delta_{\boldsymbol{\beta}}/h_n + \frac{1}{2} h_n^2 \tilde{\mathbf{X}}_i^\top \ddot{\mathbf{f}}_0(\mathbf{x})(\mathbf{X}_{i\mathbf{x}}^\top \boldsymbol{\beta}/h_n)^2 \\
& \quad + \frac{1}{2} \tilde{\mathbf{X}}_i^\top \ddot{\mathbf{f}}_{n0}(\mathbf{x}) \delta_{\boldsymbol{\beta}}^\top \mathbf{X}_{i\mathbf{x}} \mathbf{X}_{i\mathbf{x}}^\top \delta_{\boldsymbol{\beta}} + h_n \tilde{\mathbf{X}}_i^\top \ddot{\mathbf{f}}_{n0}(\mathbf{x})(\mathbf{X}_{i\mathbf{x}}^\top \boldsymbol{\beta}/h_n) \mathbf{X}_{i\mathbf{x}}^\top \delta_{\boldsymbol{\beta}} \\
& \quad + \mathcal{O}(|\delta_{\boldsymbol{\beta}}|^3 + h_n^3 | \mathscr{D},\ \Theta_n),
\end{aligned}
$$

which holds for $\mathbf{X}_i$ in close proximity of a given $\mathbf{x} \in \mathscr{D}$. ■

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Int. Symp. Information Theory*. Budapest: Akademiai Kiado.

Antoniadis, A. and Fan, J. (2001) Regularization of Wavelet Approximations. *Journal of the American Statistical Association*, **96**, 939-967

Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, **37**, 1705-1732.

Boyd, S. and Vandenberghe, L (2004). Convex optimization. Cambridge University Press.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373-384.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, **24**, 2350-2383.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, **95**, 941-956.

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

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Local estimating equations. *Journal of the American Statistical Association*, **93**, 214-227.

Cheng, M.-Y., Zhang, W. and Chen, L.-H. (2009). Statistical estimation in generalized multiparameter likelihood models. *Journal of the American Statistical Association*, **104**, 1179–1191.

Cheng, M.-Y., Honda, T., Li, J., and Peng, H. (2014). Nonparametric independence screening and structure identification for ultra high dimensional longitudinal data. *The Annals of Statistics*, **42**, 1819–1849.

Chen, R. and Tsay, R. S. (1993). Nonlinear Additive ARX Models. *Journal of the American Statistical Association*, **88**, 955-967.

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

Fan, J. and Gijbels, I. (1996). Local Polynomial Modeling and Its Applications, London: Chapman & Hall.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, **27**, 1491-1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* , **27**, 715-731.

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., Zhang, C. and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *The Annals of Statistics*, **29**, 153-193.

Fan, J., Yao, Q. and Cai, Z. (2003). Adaptive varying-coefficient linear models. *Journal of Royal Statistical Society, Series B*, **65**, 57-80.

Fan, J. and Huang, t. (2003). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, **11**, 1031-1057.

Fan, J. and Li, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association*, **99**, 710-723.

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. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, **1**, 179–195.

Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics*, **3**, 521-541.

Fan, J., Lv, J. and Qi, L., (2011). Sparse high dimensional models in economics. *Annual Review of Economics*, **3**, 291-317.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of Royal Statistical Society, Series B*, **75**, 531-552.

Fan, Y., Kong, Y., Li, D. and Zheng, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics*, **43**, 1243-1272.

Fan, Y. and Lv, J. (2016). Innovated scalable efficient estimation in ultra-large Gaussian graphical models. *The Annals of Statistics*, **44**, 2098-2126.

Fang, X., Li, J., Wong, W. K., and Fu, B. (2014). Detecting the violation of variance homogeneity in mixed models. *Statistical Methods in Medical Research*, 0962280214526194.

Frank, E. and Friedman, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135.

Hardle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, **84**, 986–995.

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

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

Huang, J., and Xie, H. (2007). Asymptotic oracle properties of SCAD-penalized least squares estimators. *Lecture Notes-Monograph Series*, 149–166.

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

Kauermann, G. and Tutz, G. (1999). On model diagnostics using varying coefficient models. *Biometrika* **86**, 119–128.

Kong, E. and Xia, Y. (2006). Variable selection for the single-index model. *Biometrika*, **94**, 217-229.

Lavergne, P. (1998). A Cauchy-Schwarz inequality for expectation of matrices.

Li, R. and Liang, H. (2008). Variable Selection in Semiparametric Regression Modeling. *The Annals of Statistics*, **36**, 261–286.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates.

Journal of the American Statistical Association, **109**, 266–274.

Li, D., Ke, Y. and Zhang, W. (2015). Model selection and structure specification in ultra-high dimensional generalised semi-varying coefficient models. *The Annals of Statistics*, **43**, 2676-2705.

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics*, **12**, 758–765.

Nolan, D. and Pollard, D. (1987). U-processes: Rates of convergence. *The Annals of Statistics*, **15**, 780-799.

Pakes, A. and Pollard, D. (1989). Simulation and the Asymptotics of Optimization Estimators. *Econometrica* **57**, 1027-1057.

Pollard, D. (1984). Convergence of Stochastic Processes.

Ruppert, D. and Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, , 22, 1346-1370.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7**, 221–264.

Song, R., Yi, F. and Zuo, H. (2012). On varying-coefficient independence screening for high-dimensional varying-coefficient models. *Statistica Sinica*, **24**, 1735–1752.

Stefanski, L. A., Wu, Y., and White, K. (2014). Variable selection in nonparametric classification via measurement error model selection likelihoods. *Journal of the American Statistical Association*, **109**, 574-589.

Schmidt, G., Mattern, R., and Schüler, F. (1981). Biomechanical investigation to determine physical and traumatological differentiation criteria for the maximum load capacity of head and vertebral column with and without protective helmet under the effects of impact. *EEC Research Program on Biomechanics of Impacts. Final report Phase III, Project 65, Institut für Rechtsmedizin, Universität Heidelberg, Germany*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the American Statistical Association, Series B*, **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, **16**, 385-395.

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

Wang, H., Li, R. and Tsai. C.-T. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, **94**, 553-568.

Wang, H. and Xia, Y. (2009). Shrinkage Estimation of the Varying Coefficient Model. *Journal of the American Statistical Association*, **104**, 747-752.

Wang, L., Peng, B., and Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *Journal of the American Statistical Association*, **110**, 1658-1669.

Wu, C. O., Chiang, C. -T. and Hoover, D. R. (1998). Asymptotic Confidence Regions for Kernel Smoothing of a Varying-Coefficient Model with Longitudinal Data. *Journal of the American Statistical Association*, **93**, 1388-1402.

Xia, Y. and Li, W. K. (1999) On single-index coefficient regression models. *Journal of the American Statistical Association*, **94**, 1275-1285.

Xia, Y. (2006) Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**, 1112-1137.

Yang, Y. (2005). Can the strengths of aic and bic be shared? A conflict between model identification and regression estimation. *Biometrika*, **92**, 937–950.

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

Zhang, X., Wu, Y., Wang, L., and Li, R. (2016). Variable Selection for Support Vector Machines in Moderately High Dimensions. *Journal of the Royal Statistical Society, Series B*, **78**, 53-76.

Lin, Y. and Zhang, H. H. (2003). Component selection and smoothing spline analysis of variance models. *The Annals of Statistics*, **34**, 2272-2297.

Zhang, W.and Lee, s.-Y. (2000). Variable Bandwidth Selection in Varying-Coefficient Models. *Journal of Multivariate Analysis*, **74**, 116-134.

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

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