HOMOGENEITY PURSUIT AND STRUCTURE IDENTIFICATION IN FUNCTIONAL-COEFFICIENT MODELS

LINGLING WEI

PhD

UNIVERSITY OF YORK

MATHEMATICS

SEPTEMBER 2018

Abstract

This thesis explores the homogeneity of coefficient functions in nonlinear models with functional coefficients, and identifies the semiparametric modelling structure. With initial kernel estimate of each coefficient function, we combine the classic hierarchical clustering method and a generalised version of the information criterion to estimate the number of clusters each of which has the common functional coefficient and determine the indices within each cluster. To specify the semi-varying coefficient modelling framework, we further introduce a penalised local least squares method to determine zero coefficient, non-zero constant coefficients and functional coefficients varying with the index variable. Through the nonparametric kernel-based cluster analysis and the penalised approach, the number of unknown parametric and nonparametric components in the models can be substantially reduced and the aim of dimension reduction can be achieved. Under some regularity conditions, we establish the asymptotic properties for the proposed methods

such as consistency of the homogeneity pursuit and sparsity. Some numerical studies including simulation and two empirical applications are given to examine the finite-sample performance of our methods.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to sincerely gratitude to my supervisors and all the academic staff in our statistics and probability group in Mathematics Department. Professor Wenyang Zhang, Professor Degui Li and Dr Marina Knight, they are outstanding statisticians and I am very grateful for their continuous support, inspiration and guidance to my Ph.D study and research. I would like to give thanks to all their time and patience they spent on training my research abilities during my research period. Its my pleasure to join in this world-class research team.

Secondly, I would also like to give thanks to all the admin staff at the Department of Mathematics at the University of York, and in particular Mr Nicholas Page , Mrs Linda Elvin, Mrs Claire Farrar. Thanks for their continuous help and support over last a few years.

I thank my Ph.D friends Mr. Jiraroj Tosasukul, Dr Xiaocheng Kou, Ms Zhongmei Ji, Dr.Yuan Ke and Dr John Box for all the support, fascinating

discussions and all their advice throughout this project.

Finally, I want to say thanks to my family: my father Shouping Wei, my mother Caixiu Wu and my devoted husband Jialong Lu, for their love and support.

Declaration

The literature review in Chapter 2 summarises some key ideas related to this thesis. In particular:

- Section 2.1 contains a review of the fundamental concepts in local polynomial modelling and is based on a summary of the first three chapters of the book *Local polynomial modelling and its applications* by Fan and Gijbels (1996).

- Section 2.2 reviews literature concerning varying coefficient models found in Fan and Zhang (1999), Fan and Zhang (2008).

- Section 2.3 is a review of extensions of the varing coefficient model and variable selection methods, which are mainly based on (but not limited to) the following literature: Cai, Fan, and Li (2000); Xia, Zhang and Tong (2004); Fan and Huang (2005).

- Section 2.4 contains a review of variable selection methodology, which is mainly based on (but not limited to) the following literature: Wang and Xia (2009), Ke, Fan and Wu (2015), Vogt and Linton (2017).

The remaining chapters are main part of my submitted paper: *Nonparametric homogeneity pursuit in functional-coefficient models*, joint with Prof. Degui Li and Prof. Wenyang Zhang. The paper is under review by The Journal of Royal Statistical Society: Series B.

To the best of my knowledge and belief this thesis does not infrigne the copyright of any other person. 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.

Chapter 1

Introduction

Regression modelling is one of the most important topics in statistical data analysis and has wide applications in various disciplines such as economics, finance and genetics. It is well know that the parametric linear model defined by

$$Y_t = \mathbf{X}_t^{\intercal}\boldsymbol{\beta}_0 + \varepsilon_t, \quad t = 1, \cdots, n, \tag{1.1}$$

has played a dominant role in regression analysis. Here Y_t is a response variable, $\mathbf{X}_t = (X_{t1}, \cdots, X_{tp})^{\intercal}$ is a p-dimensional vector of random covariates, $\boldsymbol{\beta}_0 = \left[\beta_1^0, \cdots, \beta_p^0\right]^{\intercal}$ is a p-dimensional vector of coefficients and ε_t is an independent and identically distributed (*i.i.d.*) error term. From (1.1), the linear regression relationship between Y_t and X_t is determined by the parameter vector β_0 The unknown parameter vector β_0 can be consistently

estimated by some commonly-used methods such as ordinary least squares and maximum likelihood. However, the parametric linear model assumption is often too restrictive and may be rejected by some model specification test in real data analysis. The parametric estimation based on misspecified models would provide inaccurate regression relationship of the variables which we are interested.

Comparing with the traditional parametric linear models, nonparametric models are more flexible in capturing the regression relationship and they can avoid some restrictive pre-specified parametric assumptions. When the number of covariates is large, direct nonparametric estimation would have the "curse of dimensionality" problem which is first introduced by Bellman (1957). Due to the curse of dimensionality, the convergence of nonparametric estimation to the true smooth function becomes quite slow even when the dimension is only larger than three. Therefore, how to avoid the curse of dimensionality is an important research topic in nonparametric regression estimation (Hastie and Tibshirani 1993; Fan, Yao and Cai 2003).

The main focus of this thesis is the so-called functional-coefficient model, which is an important member of nonparametric regression family and avoids the curse of dimensionality. The functional-coefficient model is a natural extension of the classic linear regression model (1.1) by allowing

the regression coefficients to vary with certain index variable, and can thus capture flexible dynamic relationship between the response and covariates. The functional-coefficient model is defined by

$$Y_t = \mathbf{X}_t^{\intercal}\boldsymbol{\beta}_0(U_t) + \varepsilon_t, \quad t = 1, \cdots, n, \tag{1.2}$$

where Y_t, $\mathbf{X}_t$ and ε_t are defined as those in model (1.1), $\boldsymbol{\beta}_0(\cdot) = \left[\beta_1^0(\cdot), \cdots, \beta_p^0(\cdot)\right]^{\intercal}$ is a p-dimensional vector of functional coefficients and U_t is a univariate index variable. In recent years, there have been extensive studies on estimation and model selection for the functional-coefficient model (1.2) and its various generalised versions, see, for example, Chen and Tsay (1993); Hastie and Tibshirani (1993); Fan and Zhang (1999; 2008), Cai, Fan and Yao (2000), Xia, Zhang and Tong (2004), Wang and Xia (2009), Kai, Li and Zou (2011), Park et al. (2015) and the references therein.

However, when the number of functional coefficients is large or moderately large, it is well-known that a direct nonparametric estimation of the potentially p different coefficient functions in model (1.2) would be very unstable. To address this problem, there have been some extensive studies in the literature on either selecting significant variables in functional-coefficient models (Fan, Ma and Dai 2014; Liu, Li and Wu, 2014), or exploring certain rank-reduced structure in functional coefficients (Jiang et al. 2013; Chen,

Li and Xia 2018), both of which aims to reduce the dimension of unknown functional coefficients and improve model estimation efficiency.

In this thesis, we consider a different approach and impose a homogeneity structure on model (1.2), i.e., the individual functional coefficients can be grouped into a number of clusters and the coefficients have the same functional pattern within each cluster. We allow that the dimension p may depend on the sample size n and can be divergent with n, but the number of unknown clusters is assumed to be fixed and much smaller than p. It is easy to see that the dimension reduction through homogeneity pursuit is more general than the commonly-used sparsity assumption in high-dimensional functional-coefficient models (c.f., Fan, Ma and Dai, 2014; Liu, Li and Wu 2014; Li, Ke and Zhang 2015) as the latter can be seen as a special case of the former with a very large group of zero coefficient. Specifically, we assume the following homogeneity structure on model (1.2): there exists a partition of $\{1, 2, \cdots, p\}$ denoted as $\mathcal{C}_0 = \left\{\mathcal{C}_1^0, \cdots, \mathcal{C}_{K_0}^0\right\}$ such that

$$\beta_j^0(\cdot) = \alpha_k^0(\cdot) \ \text{ for } \ j \in \mathcal{C}_k^0, \ \ \mathcal{C}_{k_1}^0 \cap \mathcal{C}_{k_2}^0 = \emptyset \ \text{ for } \ 1 \le k_1 \ne k_2 \le K_0, \quad (1.3)$$

where the Lebesgue measure of $\left\{u \in \mathcal{U} : \alpha_{k_1}^0(u) - \alpha_{k_2}^0(u) \ne 0\right\}$ is positive and bounded away from zero for $1 \le k_1 \ne k_2 \le K_0$, and $\mathcal{U}$ is a compact support of the index variable U_t. Furthermore, some of the functional

coefficients $\alpha_k^0(\cdot)$ are allowed to have constant values including the value of zero, indicating that model (1.2) is semi-parametric with a combination of constant and functional coefficients. Our main interests are

- explore the homogeneity structure (1.3) by estimating the *unknown* clusters $\mathcal{C}_1^0, \cdots, \mathcal{C}_{K_0}^0$ and the *unknown* number of clusters;
- identify the clusters of constant coefficients and those of coefficients varying with the index variable U_t and estimate the *unknown* components in each cluster.

The topic investigated in this thesis has two close relatives in the existing literature. On one hand, the functional-coefficient regression with the homogeneity structure is a natural extension of the linear regression with the homogeneity structure and the latter has received increasing attention in recent years. For example, Tibshirani et al.(2005) introduce the so-called fused LASSO method to study the slope homogeneity; Bondell and Reich (2008) propose the OSCAR penalised method for grouping pursuit; Shen and Huang (2010) use a truncated L_1 penalised method to extract the latent grouping structure; and Ke, Fan and Wu (2015) propose the CARDS method to identify the homogeneity structure and estimate the parameters simultaneously. On the other hand, our topic is also relevant to some recent literature on longitudinal/panel data model classification. For example, Ke,

Li and Zhang (2016) and Su, Shi and Phillips (2016) consider identifying the latent group structure for linear longitudinal data models by using the binary segmentation and shrinkage method, respectively; Su, Wang and Jin (2017) propose a penalised sieve estimation method to identify latent grouping structure for time-varying coefficient longitudinal data models and Vogt and Linton (2017) introduce a kernel-based classification of univariate nonparametric regression functions in longitudinal data models. The methodology of nonparametric homogeneity pursuit developed in this thesis will be substantially different from those in the aforementioned literature.

In this thesis, we first estimate each functional coefficient in model (1.2) by using the kernel smoothing method and ignoring the homogeneity structure (1.3), and calculate the L_1-distance matrix between the estimated functional coefficients. Then, we combine the classic hierarchical clustering method and a generalised version of the information criterion to explore the homogeneity structure (1.3), i.e., estimate K_0 and the members of $\mathcal{C}_k^0$, $k = 1, \cdots, K_0$. Under some mild conditions, we show that the developed estimators for the number K_0 and the index sets $\mathcal{C}_k^0$, $k = 1, \cdots, K_0$, are consistent. After estimating the structure (1.3), we further specify the semi-varying coefficient modelling framework by determining the zero coefficient, non-zero constant coefficients and functional coefficients varying with the index variable. This

is done by using a penalised local least squares method, where the penalty function is the weighted LASSO with the weights defined via the derivative of the well-known SCAD penalty introduced by Fan and Li (2001). With the nonparametric cluster analysis and the penalised approach, we may find that the number of the unknown components in model (1.2) can be reduced from p to $K_0 - 1$ (if the zero constant exists in the model). Consequently, we achieve the aim of dimension reduction in the functional-coefficient model.

In addition, the choice of the tuning parameters in the proposed estimation approach is discussed and the relevant computational algorithm is introduced. The simulation studies show that the proposed methods have reliable finite-sample numerical performance. We finally apply the model and methodology to analyse the Boston house price data as well as the plasma beta-carotene level data, and find that the original nonparametric functional-coefficient models can be simplified and the number of unknown components involved can be substantially reduced. In particular, the out-sample mean squared prediction errors using our approach are usually much smaller than those using the naive kernel method which ignores the latent homogeneity structure. The rest of the thesis is organised as follows.

- Chapter 2: We briefly review the existing models and methods which are relevant to the proposed method. They include kernel-based

nonparametric estimation, functional-coefficient models and their extensions high-dimensional variable selection methods.

- Chapter 3: We introduce the kernel-based hierarchical clustering method and a generalised information criterion to estimate the homogeneity structure specified in (1.3). Furthermore, a penalised method is proposed to determine zero-coefficient, non-zero constant-coefficients and functional-coefficients in the model. The computational algorithm is to implement the proposed methods and the choice of the tuning parameters are also given in this chapter.

- Chapter 4: We report two Monte-Carlo simulation studies and two real data applications (Boston house data and plasma beta-carotene level data separately) to evaluate the finite-sample performance of the proposed methodology.

- Chapter 5: We establish the asymptotic theory for the proposed clustering and estimation methods. The detailed proofs of the main asymptotic theorems are shown as well.

- Chapter 6: We conclude the thesis and discuss some possible extensions.

Chapter 2

Literature Review

In this chapter, we first review the univariate nonparametric modelling framework and local polynomial estimation approach which are fundamental tools of our research in this thesis. In Section 2.2, the varying-coefficient model and its nonparametric estimation methods are introduced. Section 2.3 gives some extensions of the conventional varying-coefficient model. Section 2.4 contains a brief review of the variable selection for both the linear models and varying-coefficient models. Different penalised methods including LASSO, SCAD, KLASSO are discussed in Section 2.4. Homogeneity pursuit in linear regression models will be briefly reviewed in this section as well.

2.1 Univariate Nonparametric Modelling

As introduced in Chapter 1, the parametric linear and nonlinear model need some pre-specified parametric assumptions before developing feasible estimation. For example, the collected data are often assumed to follow some distribution with parameters to be estimated. However, in practical applications, the parametric model assumptions are often rejected by real data, leading to rapid development of distribution free models and data-driven nonparametric methods in recent decades. The latter neither requires the data set to follow a specific distribution nor assumes a parametric form on regression functions. It allows data to "speak for themselves" when determining the functional form. The nonparametric modelling methods have been applied in a wide range of disciplines including biology, economics and public health. In this section, we review the nonparametric regression model with univariate regressor (to avoid the curse of dimensionality) and introduce the kernel-based local polynomial estimation method which is systematically studied by Fan and Gijbels (1996).

Consider an independent and identically distributed bivariate data sample $(X_1, Y_1) \cdots (X_n, Y_n)$ collected from the population (X, Y). The nonpara-

metric regression model is defined by

$$Y = m(X) + \sigma(X)\epsilon, \tag{2.1}$$

where X and ϵ are assumed to be independent for simplicity, $\mathrm{E}(\epsilon) = 0$, $\mathrm{Var}(\epsilon) = 1$, $m(\cdot)$ is the mean regression function and $\sigma^2(X)$ is the variance function. In particular, the regression function $m(\cdot)$ can be written as the conditional expectation of Y with X given, i.e., $m(\cdot) = E(Y|X = x_0)$. Our aim is to estimate the unknown mean regression function $m(x_0)$ and its derivatives $m^{(j)}(x_0), j = 1, \ldots, p$, where p is a finite positive integer.

Assume that $m(\cdot)$ has continuous derivative up to the $(p+1)^{th}$ order of derivative exists. We apply the Taylor expansion for the unknown mean regression function $m(x)$ in its neighbourhood of x_0, and approximate it by a local p-order polynomial as

$$m(x) \approx m(x_0) + m^{(1)}(x_0)(x - x_0) + \frac{m^{(2)}(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p. \tag{2.2}$$

We may treat $m(x_0), m^{(1)}(x_0), m^{(2)}(x_0), \cdots, m^{(p)}(x_0)$ as unknown "local parameters" to be estimated and let $\beta_j = \frac{m^{(j)}(x_0)}{j!}$, $j = 0, 1, \cdots, p$. Then we

can rewrite the local polynomial approximation (2.2) as

$$m(x) \approx \sum_{j=0}^{p} \beta_j (X_i - x_0)^j. \tag{2.3}$$

We denote $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ as estimators of $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$, which are obtained by minimising following weighted least squares objective function,

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 K_h(X_i - x_0), \tag{2.4}$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth. which is used to allocate weights to each data point. The weight in (2.4) are determined by $K(\cdot)$ and h. We will discuss choice of the kernel function and bandwidth later in this section.

Denote

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

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

and $\boldsymbol{W}$ is an $n \times n$ diagonal matrix with diagonal elements being $K_h(X_i - X_0)$,

$$\boldsymbol{W} = \text{diag}(K_h(X_1 - x_0), \cdots, K_h(X_n - x_0)).$$

With the above notation, we can rewrite the weighted least squares problem (2.4) in a matrix form

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}} \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}),$$

and its solution is given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{Y}, \tag{2.5}$$

with $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \cdots, \hat{\beta}_p)^\intercal$.

Commonly-used kernel functions include: Gaussian kernel, Epanechnikov kernel and Uniform kernel, which are defined as follows.

1. *Gaussian kernel:*

$$K(x) = \frac{1}{\sqrt{2\pi}} exp(-x^2/2)$$

2. *Epanechnikov kernel:*

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2), & |x| \leq 1 \\ 0, & |x| > 1 \end{cases}$$

3. *Uniform kernel:*

$$K(x) = \begin{cases} \frac{1}{2}, & |x| \leq 1 \\ 0, & |x| > 1 \end{cases}$$

Note that when the Epanechnikov or Uniform kernel is used, to estimate the regression function at the point of x, we discard the sample data whose X observations are either larger than $x + h$ or smaller than $x - h$. Throughout the numerical studies in this thesis, the Epanechnikov kernel is used due to to its desirable statistical properties, see Fan (1992) for details.

The bandwidth selection is crucial to the local polynomial estimation as it determines the nonparametric model complexity. In numerical studies, the choice of an appropriate bandwidth plays a more important role in kernel-based estimation than the choice of kernel function. When the bandwidth value is too small, it would lead to overfitted model and result

in undersmoothed functional estimation. On the other hand, when the bandwidth value is too large, it would lead to oversmoothed nonparametric estimation, affecting approximation accuracy. Various bandwidth selection methods are available in the literature, see, for example, Ruppert, Sheather and Wang (1995), Fan and Gijbels (1996). Among them the leave-one-out cross validation (CV) method is probably the most frequently-used one (Stone, 1974). Recall objective function (2.1), the CV bandwidth selection criterion is described as follows.

$$\mathsf{CV}(h) = \sum_{i=1}^{n} \{Y_i - \hat{m}_{-i}(X_i)\}^2 w(X_i), \tag{2.6}$$

where $w(X_i)$ is a weighting function, $\hat{m}_{-i}(X_i)$ is the estimation of unknown regression function in model (2.1) and the i-th observation is removed from the sample in the estimation. There are different methods to estimate $\hat{m}_{-i}(X_i)$, eg. Nadaraya-Watson estimator, Gasser-Müller estimator. Then the optimal bandwidth can be obtained by minimizing the CV function in (2.6).

2.2 Varying-Coefficient Models

Varying coefficient models are very useful and important models, capturing flexible dynamic relationship between the covariates and response. They are a natural extension of classic linear regression models by allowing its regression coefficients to vary with certain important index variable. The varying-coefficient models have been widely used in nonlinear time series analysis and longitudinal data modelling see, for example, Chen and Tsay (1993), Hastie and Tibshirani (1993), Cai, Fan and Yao (2000). The varying-coefficient model is defined by

$$Y = \sum_{j=1}^{p} \beta_j(U) X_j + \varepsilon \tag{2.7}$$

with

$$E(\epsilon | U, X_1, \cdots, X_p) = 0 \quad a.s.,$$

and

$$Var(\epsilon | U, X_1, \cdots, X_p) = \sigma^2(U) \quad a.s.,$$

where Y is a response variable, $X = (X_1, \cdots, X_p)^\intercal$ is a p-dimensional vector of random covariates and $\beta_j(\cdot) = \left[\beta_1^0(\cdot), \cdots, \beta_p^0(\cdot)\right]^\intercal$ is a p-dimensional vector of unknown functional coefficients, U is a univariate index variable and ε is

an independent and identically distributed (*i.i.d.*) error term. By setting the first random covariate to one ($X_1 \equiv 1$), the intercept function can be included in the model (2.7).

To estimate the functional coefficient $\beta_j(\cdot)$ in model (2.7), the local linear method (Fan 1993 ; Fan and Gijbels 1996) can be applied directly. We start with the so-called one-step estimation method. Assume that the coefficient functions have continuous derivatives up to the second order. Then, we approximate the coefficient function locally at each point u_0 by using the Taylor expansion

$$\beta_j(u) \approx \beta_j + \dot{\beta}_j(u - u_0), \tag{2.8}$$

where u is in a small neighbourhood of u_0. Then we can obtain $\hat{\beta}(u_0)$ by using local least square method to minimize

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{p} \left\{ \beta_j + \dot{\beta}_j(U_i - u_0) \right\} X_{ij} \right\}^2 K_h(U_i - u_0), \tag{2.9}$$

This idea was proposed by Cleveland et al.(1991) but it require the degrees of smoothness are same for functions $\beta_j(u_0)$. To estimate the same degree smoothness of functional coefficients of a varying-coefficient model, above simple local regression can be used. However, the optimal estimation cannot be obtained by using one-step method if different coefficient functions have

different smoothness. Therefore, Fan and Zhang (1999) proposed a new two-step method, which may repair the weakness of the one-step method and obtain the optimal rate of estimation.

The main development of two-step method is that we can estimate coefficient functions more accurately even if different coefficient functions have different degree of smoothness (without assumption of same degree smoothness) and optimal rates of convergence can be achieved. To show the nice properties of two-step method, we make comparation between traditional one-step method and developed two-step method.

We assume $\beta_p(\cdot)$ is smoother than any $\beta_j(\cdot), j = 1, \cdots, p-1$ (same smooth degree) and has fourth derivative, where p is dimension of covariates X in the model. For one-step method, similar to model (2.7), we re-define the varying-coefficient model basd on different smoothness of functional coefficients

$$Y = \sum_{j=1}^{p-1} \beta_j(U) X_j + \beta_p(U) X_p + \varepsilon. \tag{2.10}$$

For each given u_0, we approximate the function $\beta_p(\cdot)$ locally by following cubic function

$$\beta_p(u) \approx \beta_p + \beta_p^{(1)}(u - u_0) + {\beta^{(2)}}_p(u - u_0)^2 + {\beta^{(3)}}_p(u - u_0)^3. \tag{2.11}$$

Then the one-step estimator $\hat{\beta}_p^{OS}(u_0)$ can be estimated by minimizing

$$\sum_{i=1}^{n}\Bigg\{Y_i-\sum_{j=1}^{p-1}\Big\{\beta_j+\dot{\beta}_j(U_i-u_0)\Big\}X_{ij}-$$

$$\Big\{\beta_p+\beta_p^{(1)}(U_i-u_0)+\beta^{(2)}{}_p(U_i-u_0)^2+\beta^{(3)}{}_p(U_i-u_0)^3\Big\}X_{ip}\Big\}^2\times K_{h_1}(U_i-u_0). \tag{2.12}$$

For two-step method, in first step, we first obtain the preliminary estimator. Let $\hat{\beta}_{1,0}(u_0),\cdots,\hat{\beta}_{p,0}(u_0)$ denote the initial estimate of $\beta_1(u_0),\cdots,\beta_p(u_0)$. Given initial smaller bandwidth h_0, a preliminary estimate can be obtained by minimizing

$$\sum_{i=1}^{n}\Bigg\{Y_i-\sum_{j=1}^{p}\Big\{\beta_j+\dot{\beta}_j(U_i-u_0)\Big\}X_{ij}\Bigg\}^2K_{h_0}(U_i-u_0). \tag{2.13}$$

In the second step, we substitute the preliminary estimates $\hat{\beta}_{1,0}(\cdot),\cdots,\hat{\beta}_{p-1,0}(\cdot)$ and estimate $\beta_p(u_0)$ by minimizing

$$\sum_{i=1}^{n}\Bigg\{Y_i-\sum_{j=1}^{p-1}\hat{\beta}_{j,0}(U_i)X_{ij}-$$

$$\Big\{\beta_p+\beta_p^{(1)}(U_i-u_0)+\beta^{(2)}{}_p(U_i-u_0)^2+\beta^{(3)}{}_p(U_i-u_0)^3\Big\}X_{ip}\Big\}^2\times K_{h_2}(U_i-u_0), \tag{2.14}$$

where h_2 is the bandwidth in this second step of two-step method. Through the above method, two-step estimator $\hat{\beta}_p^{TS}(u_0)$ can be obtained. Note that the initial bandwidth h_0 is small enough to reduce the bias, therefore, the

choice of h_2 is not too sensitive to the two-step estimation. As the problem in the second step is a univariate smoothing problem, the bandwidth h_2 for the second step can be selected through existing bandwidth selection procedures which we introduced in previous section. Above is the basic procedure and comparation of one-step and two-step method.

2.3 Extentions of Varying-Coefficient Model

Apart from the varying-coefficient model we reviewed, researchers may have interests in resolving the problem of semi-varying coefficient model (Cai, Fan and Li 2000; Xia, Zhang and Tong 2004; Fan and Huang 2005). That is when some of the coefficients of the varying-coefficient model are not really varying. For instance, there might be a situation that some coefficients are constant. In this section, we will brief review the semi-varying coefficient model. We define a semi-varying coefficient regression model

$$Y = \mathbf{X}^{\intercal}\boldsymbol{\beta}(U) + \boldsymbol{Z}^{\intercal}\boldsymbol{\beta}^{\star} + \varepsilon, \tag{2.15}$$

where Y is the response variable and $\{U, \mathbf{X}, \mathbf{Z}\}$ are covariates of Y, $\boldsymbol{\beta}(\cdot) = [\beta_1(\cdot), \cdots, \beta_p(\cdot)]^{\intercal}$ is a p-dimensional vector of functional coefficients and $\boldsymbol{\beta}^{\star} = \left[\beta_1^{\star}, \cdots, \beta_q^{\star}\right]^{\intercal}$ is a q-dimensional vector of constant coefficients , U is a

univariate index variable, and ε is an independent and identically distributed (*i.i.d.*) error term. Although people may say that if we set $\beta^{\star} = 0$, and the above model becomes a standard varying-coefficient, but we should not regard it as a special case of varying-coefficient. If we do so, as a result, the variance of the estimator might be higher and that is not desirable.

One method to estimate unknown parameters is called profile least squares estimation which was introduced by Fan and Huang (2005). The above semi-varying coefficient model (2.15) can be redefined as

$$Y_i = \sum_{j=1}^{p} \beta_j(U_i) X_{ij} + \sum_{j_2=1}^{q} \beta_{j_2}^{\star} Z_{ij_2} + \varepsilon_i, i = 1, \cdots, n, \tag{2.16}$$

and

$$Y_i^* = \sum_{j=1}^{p} \beta_j(U_i) X_{ij} + \varepsilon_i, i = 1, \cdots, n, \tag{2.17}$$

where $Y_i^* = Y_i - \sum_{j=1}^{q} \beta_{j_2}^{\star} Z_{ij_2}$. The above steps transform the semi-varying coefficient model into the standard varying coefficient model. Then, the local linear estimation approach can be applied to estimate the coefficient function $\beta_j(\cdot), j = 1, \cdots, p$. Here u is still the neighbourhood of u_0 as we defined in previous varying coefficient model. Recall the local linear function

to approximate $\beta_i(u)$

$$\beta_j(u) \approx \beta_j(u_0) + \dot{\beta}_j(u_0)(u - u_0),\ j = 1, \cdots, p. \tag{2.18}$$

Weighted local least square estimation can be applied to minimize

$$\sum_{i=1}^{n} \left[Y_i^* - \sum_{j=1}^{p} \left\{ \beta_j + \dot{\beta}_j(U_i - u_0) \right\} X_{ij} \right]^2 K_h(U_i - u_0). \tag{2.19}$$

Then we can obtain the estimator by following matrix form

$$(\hat{\beta}_1(u_0), \cdots, \hat{\beta}_p(u_0), \hat{\dot{\beta}}_1(u_0), \cdots, \hat{\dot{\beta}}_p(u_0))^\intercal = \{\boldsymbol{D}^\intercal \boldsymbol{W} \boldsymbol{D}\}^{-1} \boldsymbol{D}^\intercal \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}^\star), \tag{2.20}$$

where

$$\boldsymbol{D} = \begin{pmatrix} X_1^\intercal & (U_1 - u_0)X_1^\intercal \\ \vdots & \vdots \\ X_n^\intercal & (U_n - u_0)X_n^\intercal \end{pmatrix}.$$

Moreover, we denote

$$\boldsymbol{M} = \begin{pmatrix} \mathbf{X}_1^\intercal \boldsymbol{\beta}(U_1) \\ \vdots \\ \mathbf{X}_n^\intercal \boldsymbol{\beta}(U_n) \end{pmatrix},$$

and

$$\boldsymbol{H} = \begin{pmatrix} (\boldsymbol{X}_1^{\intercal} \quad 0) \left\{\boldsymbol{D}_1{}^{\intercal}\boldsymbol{W}_1\boldsymbol{D}_1\right\}^{-1} \boldsymbol{D}_1{}^{\intercal}\boldsymbol{W}_1 \\ \vdots \\ (\boldsymbol{X}_n^{\intercal} \quad 0) \left\{\boldsymbol{D}_n{}^{\intercal}\boldsymbol{W}_n\boldsymbol{D}_n\right\}^{-1} \boldsymbol{D}_n{}^{\intercal}\boldsymbol{W}_n. \end{pmatrix}.$$

Model (2.17) can be write into matrix form

$$\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}^{\star} = \boldsymbol{M} + \boldsymbol{\varepsilon}. \tag{2.21}$$

Then we obtain

$$\hat{\boldsymbol{M}} = \boldsymbol{H}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\beta}^{\star}). \tag{2.22}$$

By substituting $\hat{\boldsymbol{M}}$ into (2.21), we may obtain

$$\boldsymbol{Y} - \boldsymbol{H}\boldsymbol{Y} = (\boldsymbol{Z} - \boldsymbol{H}\boldsymbol{Z})\boldsymbol{\beta}^{\star} + \boldsymbol{\varepsilon}. \tag{2.23}$$

Applying least squares method, the estimator $\hat{\boldsymbol{\beta}}^{\star}$ can be obtained

$$\hat{\boldsymbol{\beta}}^{\star} = \left\{\boldsymbol{Z}^{\intercal}(\boldsymbol{I} - \boldsymbol{H})^{\intercal}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}\right\}^{-1} \boldsymbol{Z}^{\intercal}(\boldsymbol{I} - \boldsymbol{H})^{\intercal}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}. \tag{2.24}$$

Apart from semi-varying coefficient model, generalized varying coefficient model is also popular to use in statistics. First of all, recall generalized linear

model (Cai, Fan and Li 2000), then we may have

$$f(y|\mathbf{u},\mathbf{x}) = \exp\left\{\frac{\theta(\mathbf{u},\mathbf{x})y - b[\theta(\mathbf{u},\mathbf{x})]}{a(\phi)} + c(y,\phi)\right\}. \tag{2.25}$$

Based on a random sample $\{U_i, \mathbf{X}_i, Y_i\}$ where $i = 1, \cdots, n$ and $a(\cdot), b(\cdot), c(\cdot,\cdot)$ are given functions (McCullagh and Nelder 1989; Fan and Gijbels 1996). Define the conditional log-likelihood function

$$\ell\left\{m(\mathbf{u},\mathbf{x}), y\right\} = \theta(\mathbf{u},\mathbf{x})y - b[\theta(\mathbf{u},\mathbf{x})]. \tag{2.26}$$

Parameter ϕ has been omitted as only the mean function is what we are interested in. By comparison with the above linear case, the generalized varying coefficient model allows coefficients to vary with covariates. Therefore, we have

$$g\left\{m(u,x)\right\} = \sum_{j=1}^{p} \beta_j(u)x_j, \tag{2.27}$$

where $g(\cdot)$ is link function, x is a p-dimensional covariate and $m(u,x)$ is the mean regression function of the response variable Y, u is a covariate index variable. Local likelihood estimation method will be used to estimate varying coefficient $\beta(\cdot)$ and we locally approximate function $\beta_j(u)$ by the

function of $\beta_j(u) \approx \beta_j + \dot{\beta}_j(u - u_0)$ and denote

$$\ell(\beta, \dot{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left[g^{-1}\left\{\sum_{j=1}^{p}(\beta_j + \dot{\beta}_j(U_i - u_0))X_{ij}\right\}, Y_i\right] K_h(U_i - u_0), \quad (2.28)$$

where $K(\cdot)$ is a kernel function and $K_h(\cdot) = K(\cdot/h)/h$. $\beta = (\beta_1, \cdots, \beta_p)^\intercal$ and $\dot{\beta} = (\dot{\beta}_1, \cdots, \dot{\beta}_p)^\intercal$. Estimator $\hat{\beta}(u_0)$ will be obtained when we maximize the local likelihood function $\ell(\beta, \dot{\beta})$, where $\hat{\beta}(u_0)$ is the estimate of $\beta_1(u_0), \cdots, \beta_p(u_0)$ and $\hat{\dot{\beta}}(u_0)$ is the estimate of $\dot{\beta}_1(u_0), \cdots, \dot{\beta}_p(u_0)$. In order to simplify notations, we denote $B(u_0) = (\beta_1, \cdots, \beta_p, \dot{\beta}_1, \cdots, \dot{\beta}_p)^\intercal$. If we use local maximum likelihood estimation to get $\hat{\beta}_j(\cdot)$, it would be quite computation consuming since we need to maximize $\ell(\beta, \dot{\beta})$ for too many distinct values of u_0 by using iterative method. In order to reduce computation cost, we may use one-step local maximum likelihood estimation method. The idea is by given an initial estimator $\hat{B}_0(u_0) = (\hat{\beta}(u_0)^\intercal, \hat{\dot{\beta}}(u_0)^\intercal)$, followed by using the one-step Newton-Raphson algorithm to find an updated estimator

$$\hat{B}_{OS}(u_0) = \hat{B}_0(u_0) - \left\{\ell''(\hat{B}_0(u_0))\right\}^{-1} \ell'(\hat{B}_0(u_0)), \quad (2.29)$$

where $\ell'(B)$ and $\ell''(B)$ are gradient and Hessian matrix of local likelihood $\ell(B)$.

Now, we can know the estimated number of group. In next Section, we

will review some existing variable selection methods.

2.4 Variable Selection Methods

As we all know, the set of variables of a well-established statistical model should be fixed and small. Redundant predictors should be removed from the model, especially for high-dimensional data. In last a few decades, penalised least squares method plays a significant role in variable selection method. Comparing with traditional model selection approaches (eg. stepwise regression), it is less computationally time consuming and quite popular to use in recent years.

The key idea of penalized least squares method is by applying a penalty function, we shrinkage some small value of coefficients to zero automatically and delete those zero coefficients in the end. Thus, we can simplify the original model. One of the most popular method for linear regression is so-called least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996). To further resolve the inconsistency issue of LASSO, Zou (2006) proposed adaptive LASSO and extension from LASSO to group LASSO was developed by Yuan and Lin (2006). Then, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty and proofed its nice properties of continuity, sparsity and unbiasedness.

Consider the linear regression model. The LASSO resolves the L_1-penalised regression problem of estimating the $\hat{\beta}_j$ to minimize

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}^{\intercal}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.30}$$

The L_1-penalty form in model (2.30) is the reason why LASSO can do shrinkage and variable selection. As we mentioned above, the weakness of LASSO is lack of continuity property. Therefore, Fan and Li (2001) proposed SCAD method and proofed its Oracle properties. In this case, the estimated estimator can be estimated as good as oracle estimator by applying SCAD penalty. We define the SCAD penalty via its derivative as

$$p'_{\lambda}(z) = \lambda \left[\mathsf{I}(z \leq \lambda) + \frac{(a_*\lambda - z)_+}{(a_* - 1)\lambda} \mathsf{I}(z > \lambda) \right], \quad a_* = 3.7.$$

Figure 2.1 shows the L_1 penalty and SCAD penalty functions separately based on the values of β. Here we design the values of β as a sequence from -5 to 5 with interval 0.1.

Apart from the above variable selection methods for linear models, Wang and Xia (2009) proposed the idea of extending the LASSO to varying coefficient models with local constant kernel estimation. The proposed method can identify the true model consistently by using the local constant

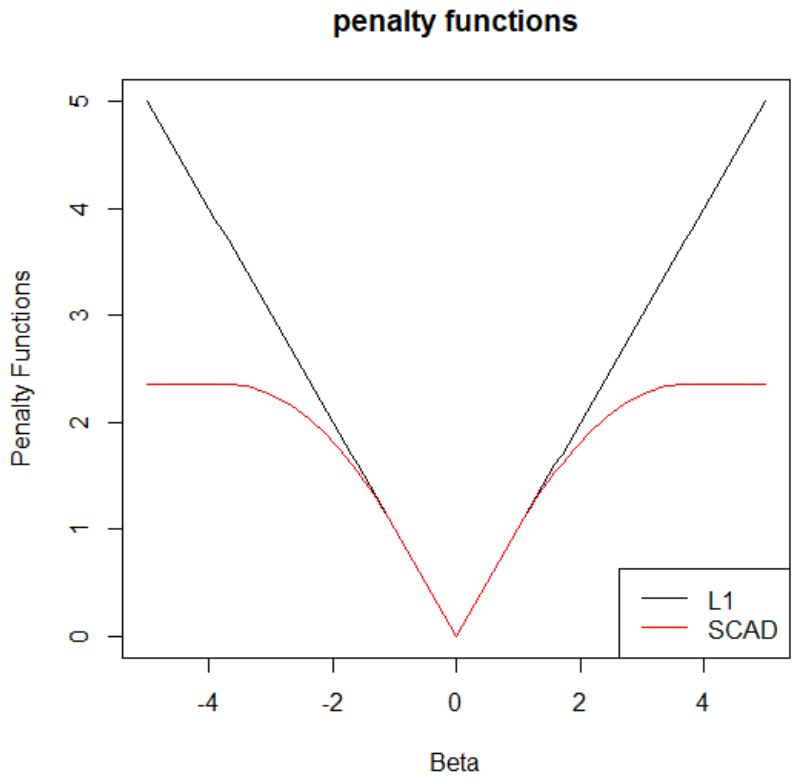

Figure 2.1: Penalty functions for L_1 (Black line)and SCAD (Red line)

estimator and the adaptive LASSO penalty. Because the method is based on the combination of kernel smoothing and LASSO, it is therefore named as Kernel LASSO (KLASSO). Followed by model (2.7), we first obtain the initial estimator $\hat{\beta}_{j,0}(\cdot)$ by locally weighted least squares function

$$\sum_{i=1}^{n}(Y_i - X_i^{\intercal}\beta(U_i))^2 K_h(U_i - u_0). \tag{2.31}$$

Then, we apply the shrinkage technique to do variable selection and propose

the following penalised estimate

$$\sum_{i=1}^{n}(Y_i - X_i^{\intercal}\beta(U_i))^2 K_h(U_i - u_0) + \sum_{j=1}^{p}\lambda_j \parallel \hat{\beta}_{j,0} \parallel . \tag{2.32}$$

Denote the minimized resulting estimator by $\hat{\beta}_{\lambda,j}(\cdot)$. Where $\| \cdot \|$ stands for Euclidean norm and $\lambda = (\lambda_1, \cdots, \lambda_p)^{\intercal} \in R^p$ is the tuning parameter. Local Quadratic Approximation will be used here and $\hat{\beta}_{j,0}(\cdot)$ is the initial estimator of the iterative algorithm. We do m^{th} iteration of KLASSO method and define the m^{th} iterative penalised estimate

$$\sum_{t=1}^{n}\sum_{i=1}^{n}(Y_i - X_i^{\intercal}\beta(U_t))^2 K_h(U_i - u) + \sum_{j=1}^{d}\lambda_j \frac{\parallel \hat{\beta}_{j,0} \parallel}{\parallel \hat{\beta}_{\lambda,j}^{(m)} \parallel}. \tag{2.33}$$

Then we may obtain

$$\hat{\beta}_{\lambda}^{(m+1)}(u_0) = [\sum_{i=1}^{n} X_i X_i^{\intercal} K_h(U_i - u_0) + D^{(m)}]^{-1} \times [\sum_{i=1}^{n} X_i X_i^{\intercal} K_h(U_i - u_0)], \tag{2.34}$$

where $D^{(m)}$ is the j^{th}diagonal component $(d \times d)$ of $\frac{\lambda_j}{\|\hat{\beta}_{\lambda,j}^{(m)}\|}$, $j = 1, \cdots, p$.

So far, we have reviewed several variable selection methods without homogeneity pursuit. It should be noticed that the concept of homogeneity has received increasing attention in recent years, for example, Tibshiranie et al.(2005); Friedman et al.(2007). Now, I will briefly review a few homogeneity

structure based methods. To start with a linear case, Ke, Fan and Wu (2015) developed a method which is called Clustering Algorithm in Regression via Data-driven Segmentation (CARDS). Define a matrix form of a linear model

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.35}$$

where $\mathbf{Y}=(y_1, \cdots, y_n)^\intercal$, $\mathbf{X} = (X_1, \cdots, X_p)$ is a $n \times p$ dimensional matrix, the parameter $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\intercal$ and $\boldsymbol{\varepsilon}$ is an *i.i.d.* error term. Note that there are two methods are introduced, which are Basic version of CARDS (bCARDS) and Advanced version of CARDS (aCARDS) separately.

For bCARDS, denote $\hat{\beta}$ be a preliminary estimator. The main idea for generating a homogeneity structure is

(i) Rearrange the coefficients $\hat{\beta}$ in ascending order.

(ii) Group the adjacent indices whose coefficients in $\hat{\beta}$ are close each other (penalised least squares method can be applied to extract the grouping structure).

(iii) In each estimated group, we force those indices to share a common coefficient and then we refit the model.

There are two steps to shrink coefficients of adjacent indices toward homogeneity structure:

Firstly, we rank the preliminary estimator in ascending order, i.e.,

$$\hat{\beta}_{(1)} \leq \hat{\beta}_{(2)} \leq \cdots \leq \hat{\beta}_{(p)}. \tag{2.36}$$

Secondly, with SCAD penalty function $P_\lambda(\cdot)$ and parameter λ, $\widetilde{\boldsymbol{\beta}}$ can be estimated by minimizing

$$\frac{1}{2n}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \sum_{j=1}^{p-1} p_\lambda(|\beta_{(j+1)} - \beta_{(j)}|). \tag{2.37}$$

For aCARDS, less information from $\widetilde{\beta}$ will be used but two penalty terms are needed. Similar with aCARDS, given a preliminary estimator $\hat{\beta}$ and get preliminary ranking $\hat{\beta}_{(1)} \leq \hat{\beta}_{(2)} \leq \cdots \leq \hat{\beta}_{(p)}$. For a tuning parameter $\delta > 0$, construct an ordered segmentation Υ where $\hat{\beta}_{(j)} - \hat{\beta}_{(j-1)} > \delta$. This is the main difference between aCARDS and bCARDS. In fact, the bCARDS is just a special case when $\delta = 0$. In the end, we compute the solution $\hat{\boldsymbol{\beta}}$ which minimizes

$$\frac{1}{2n}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\beta}), \tag{2.38}$$

where λ_1 and λ_2 are tuning parameters and $P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\beta})$ is a hybrid pairwise

penalty which is defined by

$$P_{\Upsilon,\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \sum_{l=1}^{L-1} \sum_{i\in B_l, j\in B_{l+1}} p_{\lambda_1}(|\beta_i - \beta_j|) + \sum_{l=1}^{L-1} \sum_{i,j\in B_l} p_{\lambda_2}(|\beta_i - \beta_j|), \quad (2.39)$$

where $B_1, \cdots, B_l$ is the order of segments. The first part penalty is called between-segment penalty. It penalise the pairs of indices from two adjacent segments. While the second part penalty is named within-segment penalty. Which is not rely on the ordering within the segment as it penalise all pairs of indices in each single segment.

Apart from the above methodology for homogeneity pursuit, Vogt and Linton (2017) develop a clustering method for nonparametric model with heterogeneous regression functions. Here I will briefly review the proposed classification method. The designed model is similar with the model (2.1) but here we construct a panel data model and replace ϵ by u. Specifically, we define $u_{it} = \alpha_i + \gamma_t + \epsilon_{it}$, where $i = 1, \cdots, n$ denotes the i^{th} individual, $t = 1, \cdots, T$ denotes the time point of observation, α_i is an unobserved individual, γ_t is time specific error terms which may be correlated with the regressors in an arbitrary way and ϵ_{it} is an independent and identically distributed error term. Let $g_1, \cdots, g_{k_0}$ be functions associated with these

sets. We suppose that

$$m_i = g_k, i \in C_k^0, 1 \leq k \leq K_0. \tag{2.40}$$

In order to estimate function m_i, Nadaraya-Watson, local linear or local polynomial estimator can be applied directly. Here we omit the detail of specific estimation method and focus on the proposed classification method. Assuming the true number of group is K_0, we define

$$\Delta_{ij} = \Delta_{(m_i,m_j)} = \int (m_i(u) - m_j(u))^2 f_U(u) du, \tag{2.41}$$

where $f_U(u)$ is some weight function. The classification structure can be obtained by following algorithm

Step 1: Order the distances by $\Delta_{i(1)} \leq, \cdots, \leq \Delta_{i(ns)}$, where $i \in S$ is the index and Δ_{ij} denote the weighted squared L_2-distance.

Step 2: The position of the largest jump j_{max} can be determined by ,

$$\max_{2 \leq j \leq ns} |\Delta_{1(j)} - \Delta_{1(j-1)}|.$$

Step 3: Partition original S into two subgroups, $S_<$ and $S_>$ separately.

Where

$$S_< = \{(1), \cdots, (j_{max} - 1)\} \text{ and } S_> = \{(j_{max}), \cdots, (ns)\}.$$

The first three steps can be regard as the segmentation method. Then, we iterate above algorithm:

(i) Apply Step 1 to Step 3 (from above algorithm) and set $S = \{1, \cdots, n\}$ and split it up into two subgroups $S_1 = S_<$ and $S_2 = S_>$

(ii) Design $\{S_1, \cdots, S_r\}$ as the partition of $\{1, \cdots, n\}$ from the above iteration steps. Select some group S_{l^*} from this partition for which can max $\delta_{ig} > 0$. Then we apply step 1 to step 3 and further split S_{l^*} into another subgroup $S_{l^*,<}$ and $S_{l^*,>}$.

Iterate above algorithm (K_0 - 1) times until all indices segment into K_0 groups. Thus, we finish classification of nonparametric regression functions.

In next Chapter, we will introduce our developed methods which are not covered by the reviewed literature.

Chapter 3

Homogeneity Pursuit and Algorithm

In this Chapter, we first introduce a clustering method for kernel estimated functional coefficients, followed by a generalised information criterion to determine the number of clusters in Section 3.1, and finally propose a penalised local linear estimation approach to specify the semi-varying coefficient modelling structure in Section 3.2.

3.1 Kernel-Based Cluster Method

Assuming that the coefficient functions have continuous second-order derivatives, we can use the kernel smoothing method (Wand and Jones 1994) to

obtain the preliminary estimation of $\beta_j^0(\cdot)$, $j = 1, \cdots, p$, and denote the resulting estimation by $\tilde{\beta}_j(\cdot)$. Let $\mathbb{Y}_n = (Y_1, \cdots, Y_n)^\intercal$, $\mathbb{X}_n = (\mathbf{X}_1, \cdots, \mathbf{X}_n)^\intercal$ and $\mathbb{W}_n(u) = \mathsf{diag}\{K_h(U_1, u), \cdots, K_h(U_n, u)\}$ with $K_h(U_t, u) = K((U_t - u)/h)$, where $K(\cdot)$ is a kernel function and h is a bandwidth which tends to zero as the sample size n diverges to infinity. Then the kernel estimation $\tilde{\boldsymbol{\beta}}(u_0)$ can be expressed as follows

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}(u_0) &= \left[\tilde{\beta}_1(u_0), \cdots, \tilde{\beta}_p(u_0)\right]^\intercal \\
&= \left[\sum_{t=1}^{n} \mathbf{X}_t \mathbf{X}_t^\intercal K_h(U_t, u_0)\right]^{-1} \left[\sum_{t=1}^{n} \mathbf{X}_t Y_t K_h(U_t, u_0)\right] \\
&= \left[\mathbb{X}_n^\intercal \mathbb{W}_n(u_0) \mathbb{X}_n\right]^{-1} \left[\mathbb{X}_n^\intercal \mathbb{W}_n(u_0) \mathbb{Y}_n\right], \qquad (3.1)
\end{aligned}$$

where u_0 is on the support of the index variable. Note that other commonly-used nonparametric estimation methods such as the local polynomial method (Fan and Gijbels 1996) and B-spline method (Green and Silverman 1994) are also applicable to obtain the preliminary estimates. Without loss of generality, we let $\mathcal{U} = [0, 1]$ be the compact support of the index variable U_t. Define

$$\tilde{\Delta}_{ij} = \frac{1}{n} \sum_{t=1}^{n} \left|\tilde{\beta}_i(U_t) - \tilde{\beta}_j(U_t)\right| \mathsf{I}(U_t \in \mathcal{U}_h), \qquad (3.2)$$

where $\mathsf{I}(\cdot)$ is the indicator function and $\mathcal{U}_h = [h, 1-h]$. The aim of truncating the observations outside $\mathcal{U}_h$ is to overcome the so-called boundary effect in

the kernel estimation. Noting that $h \to 0$, the set $\mathcal{U}_h$ can be sufficiently close to $\mathcal{U}$, and thus the information loss is negligible. In fact, $\tilde{\Delta}_{ij}$ can be viewed as a natural estimate of

$$\Delta_{ij}^0 = \int_{\mathcal{U}_h} \left|\beta_i^0(u) - \beta_j^0(u)\right| f_U(u)du, \tag{3.3}$$

where $f_U(\cdot)$ is the density function of U_t. Under some smoothness conditions on $\beta_i^0(\cdot)$ and $f_U(\cdot)$, we may show that

$$\Delta_{ij}^0 \to \int_{\mathcal{U}} \left|\beta_i^0(u) - \beta_j^0(u)\right| f_U(u)du, \quad n \to \infty.$$

From (1.2) and (3.3), we have $\Delta_{ij}^0 = 0$ for $i, j \in \mathcal{C}_k^0$, and $\Delta_{ij}^0 \neq 0$ for $i \in \mathcal{C}_{k_1}^0$ and $j \in \mathcal{C}_{k_2}^0$ with $k_1 \neq k_2$. Then we define a distance matrix among the functional coefficients by $\mathbf{\Delta}_0$ with the (i, j)-entry being Δ_{ij}^0, and obtain the corresponding estimated distance matrix by $\tilde{\mathbf{\Delta}}_n$ with the (i, j)-entry being $\tilde{\Delta}_{ij}$ defined in (3.2). It is obvious that both $\mathbf{\Delta}_0$ and $\tilde{\mathbf{\Delta}}_n$ are $p \times p$ symmetric matrices with the main diagonal elements being zeros.

We next use the well-known agglomerative hierarchical clustering method to explore the homogeneity among the functional coefficients. This clustering method starts with p clusters corresponding to the p functional coefficients. In each stage, a functional coefficient or a cluster of some common functional

coefficient is merged into another cluster. Then the number of clusters shrinks and we end with only one cluster. Such a clustering approach has been widely studied in the literature on cluster analysis (c.f., Everitt el al.2011; Rencher and Christensen 2012). However, to the best of our knowledge, there is virtually no work combining the agglomerative hierarchical clustering method with the kernel smoothing of functional coefficients in nonparametric homogeneity pursuit. This thesis fills in this gap. Specifically, the algorithm is described as follows, where the number of clusters K_0 is assumed to be known. Section 3.2 below will introduce an information criterion to determine the number K_0.

1. *Start with p clusters each of which contains one functional coefficient and search for the smallest distance among the off-diagonal elements of $\tilde{\mathbf{\Delta}}_n$.*

2. *Merge the two clusters with the smallest distance, and then re-calculate the distance between clusters and update the distance matrix. Here the distance between two clusters $\mathcal{A}$ and $\mathcal{B}$ is defined as the minimum distance between a point in $\mathcal{A}$ and a point in $\mathcal{B}$, which is called a single linkage (or nearest neighbour) method.*

3. *Repeat steps 1 and 2 until the number of clusters reaches K_0.*

Let $\tilde{\mathcal{C}}_1, \cdots, \tilde{\mathcal{C}}_{K_0}$ be the estimated clusters obtained via the above algorithm when the true number of clusters is known a priori. More generally, if the number of clusters is assumed to be K with $1 \leq K \leq p$, we stop the above algorithm when the number of clusters reaches K, and let $\tilde{\mathcal{C}}_{1|K}, \cdots, \tilde{\mathcal{C}}_{K|K}$ be the estimated clusters.

3.2 Penalised Local Linear Estimation

In practice, the true number of clusters is usually unknown and needs to be determined. When the number of clusters is assumed to be K, we define the kernel estimation for the functional coefficients:

$$\begin{aligned} \tilde{\boldsymbol{\alpha}}_K(u_0) &= \left[\tilde{\alpha}_{1|K}(u_0), \cdots, \tilde{\alpha}_{K|K}(u_0)\right]^{\intercal} \\ &= \left[\sum_{t=1}^{n} \tilde{\mathbf{X}}_{t,K} \tilde{\mathbf{X}}_{t,K}^{\intercal} K_h(U_t, u_0)\right]^{-1} \left[\sum_{t=1}^{n} \tilde{\mathbf{X}}_{t,K} Y_t K_h(U_t, u_0)\right], \end{aligned} \tag{3.4}$$

where

$$\tilde{\mathbf{X}}_{t,K} = \left(\tilde{X}_{t,1|K}, \cdots, \tilde{X}_{t,K|K}\right)^{\intercal} \quad \text{with} \quad \tilde{X}_{t,k|K} = \sum_{j \in \tilde{\mathcal{C}}_{k|K}} X_{tj},$$

$\tilde{\mathcal{C}}_{k|K}$ is defined as in Section 3.1. When the number K is larger than K_0, $\tilde{\boldsymbol{\alpha}}_K(\cdot)$ is still a uniformly consistent kernel estimate of the functional coefficients

(c.f., the proof of Theorem 2 in Section 5.1); but when K is smaller than K_0, the clustering approach in Section 3.1 results in a misspecified functional-coefficient model and $\tilde{\boldsymbol{\alpha}}_K(\cdot)$ can be viewed as the kernel estimate of the "quasi" functional coefficients which will be defined in (5.3) below.

We define the following objective function:

$$\mathsf{IC}(K) = \log\left[\tilde{\sigma}_n^2(K)\right] + K \cdot \left[\frac{\log(nh)}{nh}\right]^{\rho} \tag{3.5}$$

with $0 < \rho < 1$,

$$\tilde{\sigma}_n^2(K) = \frac{1}{n_h}\sum_{t=1}^{n}\left[Y_t - \tilde{\mathbf{X}}_{t,K}^{\intercal}\tilde{\boldsymbol{\alpha}}_K(U_t)\right]^2 \mathsf{I}(U_t \in \mathcal{U}_h) \quad \text{and} \quad n_h = \sum_{t=1}^{n}\mathsf{I}(U_t \in \mathcal{U}_h),$$

and determine the number of clusters through

$$\tilde{K} = \arg\min_{1\le K\le \bar{K}} \mathsf{IC}(K), \tag{3.6}$$

where $\bar{K}$ is a pre-specified finite positive integer which is larger than K_0. In practical application, $\bar{K}$ can be chosen the same as the dimension of covariates p if the latter is either fixed or moderately large. When ρ is relatively large, more clusters can be identified. When ρ is exactly 1, the number of estimated clusters would be the same as the dimension of original

covariates. Further detail and proof can be seen in Section 5.2. In our numerical studies, we make the value of $\rho = 0.5$. The penalty term in (3.5) can be replaced by $\log^{\rho-1}(nh)/nh$ when the dimension of covariates is fixed. If we choose ρ close to 1 and treat nh as the "effective" sample size, the above criterion would be similar to the classic Bayesian information criterion introduced by Park et al.1978). The latter has been extended to the nonparametric framework in recent years (c.f.,Wang and Xia 2009).

We next introduce a penalised approach to further identify the clusters with non-zero constant coefficients and the cluster with zero coefficient. For notational simplicity, we let $\tilde{\mathbf{X}}_t = \tilde{\mathbf{X}}_{t,\tilde{K}}$ and $\tilde{\boldsymbol{\alpha}}(u_0) = [\tilde{\alpha}_1(u_0), \cdots, \tilde{\alpha}_{\tilde{K}}(u_0)]^\intercal$ be defined similarly to $\tilde{\boldsymbol{\alpha}}_K(u_0)$ but with $K = \tilde{K}$. It is obvious that identifying the constant coefficients is equivalent to identifying the functional coefficients such that either their derivatives are zero or the deviation of the functional coefficients $D_k^0 = 0$ (c.f., Li, Ke and Zhang 2015), where

$$D_k^0 = \left\{ \sum_{t=1}^{n} \left[\alpha_k^0(U_t) - \bar{\alpha}_k \right]^2 \right\}^{1/2}, \quad \bar{\alpha}_k = \frac{1}{n} \sum_{s=1}^{n} \alpha_k^0(U_s).$$

In practice, we may construct the estimated deviation of the functional

coefficients by

$$\tilde{D}_k = \left\{ \sum_{t=1}^{n} \left[\tilde{\alpha}_k(U_t) - \frac{1}{n} \sum_{s=1}^{n} \tilde{\alpha}_k(U_s) \right]^2 \right\}^{1/2},$$

for $k = 1, \cdots, \tilde{K}$. Let

$$\mathbf{A} = \left(\mathbf{a}_1^\intercal, \cdots, \mathbf{a}_n^\intercal\right)^\intercal, \quad \mathbf{a}_t = (a_{t1}, \cdots, a_{t\tilde{K}})^\intercal;$$

$$\mathbf{B} = \left(\mathbf{b}_1^\intercal, \cdots, \mathbf{b}_n^\intercal\right)^\intercal, \quad \mathbf{b}_t = (b_{t1}, \cdots, b_{t\tilde{K}})^\intercal;$$

$$A_k = (a_{1k}, \cdots, a_{nk})^\intercal, \quad B_k = (b_{1k}, \cdots, b_{nk})^\intercal.$$

We define the penalised objective function as follows:

$$\mathcal{Q}_n(\mathbf{A}, \mathbf{B}) = \mathcal{L}_n(\mathbf{A}, \mathbf{B}) + \mathcal{P}_{n1}(\mathbf{A}) + \mathcal{P}_{n2}(\mathbf{B}), \tag{3.7}$$

where

$$\begin{aligned} &\mathcal{L}_n(\mathbf{A}, \mathbf{B}) = \sum_{s=1}^{n} \mathcal{L}_n(\mathbf{a}_s, \mathbf{b}_s) = \\ &\frac{1}{n} \sum_{s=1}^{n} \sum_{t=1}^{n} \left[Y_t - \tilde{\mathbf{X}}_t^\intercal \mathbf{a}_s - \tilde{\mathbf{X}}_t^\intercal \mathbf{b}_s (U_t - U_s) \right]^2 K_h(U_t, U_s), \\ &\mathcal{P}_{n1}(\mathbf{A}) = \sum_{k=1}^{\tilde{K}} p'_{\lambda_1}\left(\|\tilde{A}_k\|\right) \|A_k\|, \quad \mathcal{P}_{n2}(\mathbf{B}) = \sum_{k=1}^{\tilde{K}} p'_{\lambda_2}\left(\tilde{D}_k\right) \|hB_k\|, \end{aligned} \tag{3.8}$$

in which $\tilde{A}_k = [\tilde{\alpha}_k(U_1), \cdots, \tilde{\alpha}_k(U_n)]^\intercal$, $\|\cdot\|$ denotes the Euclidean norm,

λ_1 and λ_2 are two tuning parameters, $p'_\lambda(\cdot)$ is the derivative of the SCAD penalty function Fan and Li (2001)

$$p'_\lambda(z) = \lambda \left[\mathsf{I}(z \leq \lambda) + \frac{(a_*\lambda - z)_+}{(a_* - 1)\lambda} \mathsf{I}(z > \lambda) \right], \quad a_* = 3.7.$$

Let

$$\widehat{A}_k = [\widehat{\alpha}_k(U_1), \cdots, \widehat{\alpha}_k(U_n)]^\intercal \text{ and } \widehat{B}_k = [\widehat{\alpha}'_k(U_1), \cdots, \widehat{\alpha}'_k(U_n)]^\intercal, \ k = 1, \cdots, \tilde{K}, \tag{3.9}$$

be the minimiser of the objective function $\mathcal{Q}_n(\mathbf{A}, \mathbf{B})$. Through the penalisation, we would expect $\|\widehat{A}_k\| = 0$ when $\tilde{\mathcal{C}}_{k|\tilde{K}}$ is the estimated cluster of zero coefficient, and $\|\widehat{B}_k\| = 0$ when $\tilde{\mathcal{C}}_{k|\tilde{K}}$ is the estimated cluster of non-zero constant. Hence, if $\|\widehat{A}_k\| = 0$, the corresponding covariates are not significant and should be removed from the functional-coefficient model (1.2); and if $\|\widehat{B}_k\| = 0$, the functional coefficient has a constant value and can be consistently estimated by

$$\widehat{\alpha}_k = \frac{1}{n} \sum_{t=1}^{n} \widehat{\alpha}_k(U_t).$$

The following flowchart shows the flow of the proposed estimation process.

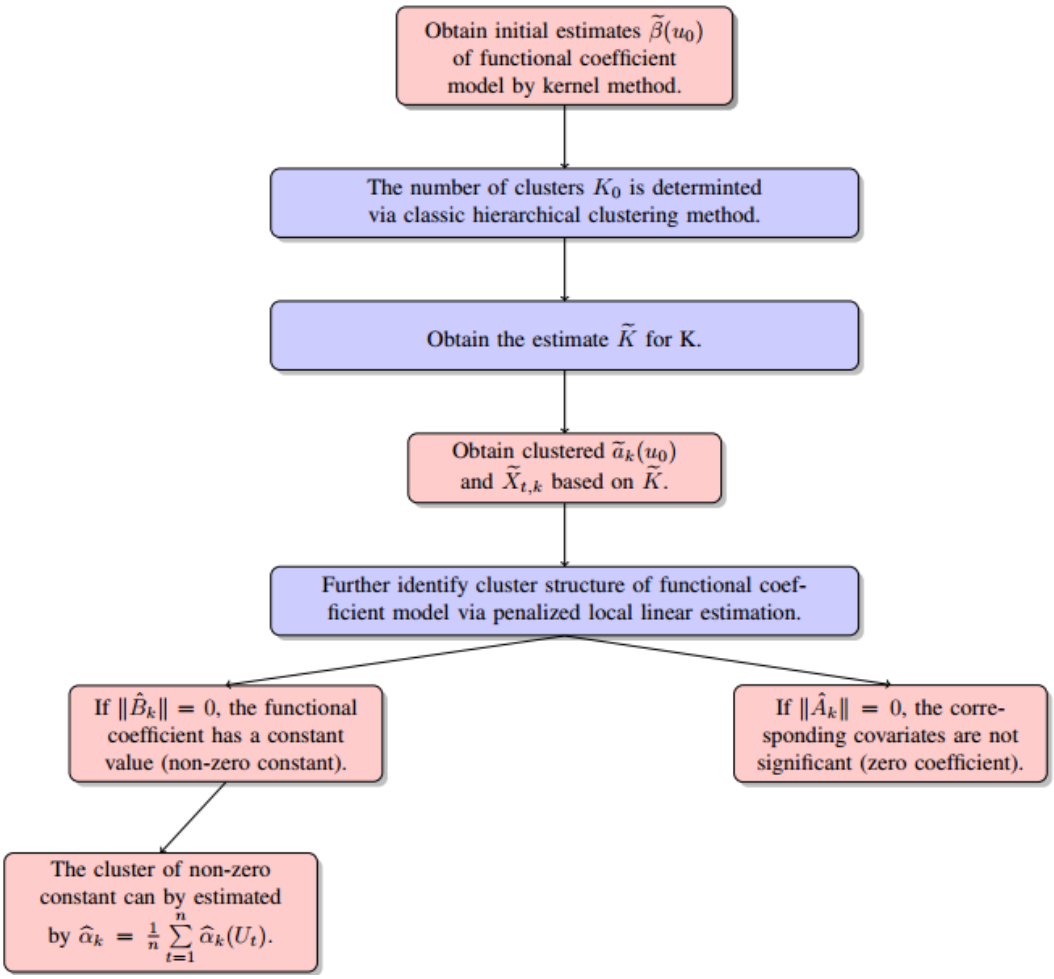

Figure 3.1: Flowchart of the proposed estimation process.

3.3 Choice of Tuning Parameters

The nonparametric kernel-based estimation may be sensitive to the value of bandwidth h. Therefore, how to choose an appropriate bandwidth is an important issue when applying our kernel-based clustering and estimation methods in practice. A commonly-used bandwidth selection method is the so-called cross-validation criterion. Specifically, the objective function for the leave-one-out cross-validation criterion is defined by

$$\mathsf{CV}(h) = \frac{1}{n}\sum_{t=1}^{n}\Big[Y_t - \mathbf{X}_t^{\intercal}\tilde{\boldsymbol{\beta}}_{-t}(U_t|h)\Big]^2, \tag{3.10}$$

where $\tilde{\boldsymbol{\beta}}_{-t}(\cdot|h)$ is the preliminary kernel estimator of $\boldsymbol{\beta}_0(\cdot)$ in model (1.2) when the bandwidth is h and the t-th observation is removed from the sample in the estimation. Then we determine the optimal bandwidth $\hat{h}_{opt}$ by minimising $\mathsf{CV}(h)$ with respect to h.

For the choice of the tuning parameters λ_1 and λ_2 in the penalised local least squares method, we use the generalised information criterion (GIC) proposed by Fan and Tang (2013), which is briefly described as follows. Let $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ and use $\mathcal{M}_1(\boldsymbol{\lambda})$ and $\mathcal{M}_2(\boldsymbol{\lambda})$ to denote the index sets of nonparametric functional coefficients and non-zero constant coefficients, respectively (after implementing the kernel-based clustering analysis and

penalised estimation with the tuning parameter vector $\boldsymbol{\lambda}$). As Cheng, Zhang and Chen (2009) suggest that an unknown functional parameter (varying with the index variable) would amount to $m_0 h^{-1}$ unknown constant parameters with $m_0 = 1.028571$ when the Epanechnikov kernel is used, we construct the following GIC objective function:

$$\begin{aligned} \mathsf{GIC}(\boldsymbol{\lambda}) &= \sum_{t=1}^{n} \left[Y_t - \sum_{k \in \mathcal{M}_1(\boldsymbol{\lambda})} \tilde{X}_{t,k|\tilde{K}} \widehat{\alpha}_{k,\lambda}(U_t) - \sum_{k \in \mathcal{M}_2(\boldsymbol{\lambda})} \tilde{X}_{t,k|\tilde{K}} \widehat{\alpha}_{k,\lambda} \right]^2 \\ &\quad + 2\ln[\ln(n)]\ln(m_0 h^{-1})(|\mathcal{M}_2(\boldsymbol{\lambda})| + |\mathcal{M}_1(\boldsymbol{\lambda})| m_0 h^{-1}), \end{aligned} \tag{3.11}$$

where $\widehat{\alpha}_{k,\lambda}(\cdot)$ and $\widehat{\alpha}_{k,\lambda}$ are defined as the penalised estimation in Section 3.2 using the tuning parameter vector $\boldsymbol{\lambda}$, $|\mathcal{M}|$ denotes the cardinality of the set $\mathcal{M}$, and the bandwidth h can be chosen as $\hat{h}_{opt}$ determined by the leave-one-out cross-validation introduced above. The optimal value of $\boldsymbol{\lambda}$ can be found by minimising the objective function $\mathsf{GIC}(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$.

3.4 Computational Algorithm

Let $\tilde{\mathbf{X}}_t = \tilde{\mathbf{X}}_{t,\tilde{K}} = \left(\tilde{X}_{t,1|\tilde{K}}, \cdots, \tilde{X}_{t,\tilde{K}|\tilde{K}} \right)^{\intercal}$ and define

$$\tilde{\boldsymbol{\Omega}}_{nk}(j) = \mathsf{diag} \left\{ \tilde{\Omega}_{nk,1}(j), \cdots, \tilde{\Omega}_{nk,n}(j) \right\}$$

with $\tilde{\Omega}_{nk,s}(j) = \frac{2}{n}\sum_{t=1}^{n} \tilde{X}_{t,k|\tilde{K}}\tilde{X}_{t,k|\tilde{K}}\left[(U_t - U_s)/h\right]^j K_h(U_t, U_s)$. It it obtained by second derivative of loss function $\mathcal{L}_n(\mathbf{A}, \mathbf{B})$ in equation (3.8). We next introduce an iterative procedure to compute the penalised local least squares estimates of the functional coefficients proposed in Section 3.2 see Li et al. (2015).

1. *Find the initial estimates of A_k^0 and B_k^0, and we denote them by*

$$\hat{A}_k^{(0)} = \left[\hat{\alpha}_k^{(0)}(U_1), \cdots, \hat{\alpha}_k^{(0)}(U_n)\right]^{\intercal} \text{ and } \hat{B}_k^{(0)} = \left[\hat{\alpha}_k^{\prime(0)}(U_1), \cdots, \hat{\alpha}_k^{\prime(0)}(U_n)\right]^{\intercal},$$

   respectively. These initial estimates can be obtained by using the conventional (unpenalised) local linear estimation method.

2. *Let $\hat{A}_k^{(j)}$ and $\hat{B}_k^{(j)}$ be the estimates after the j-th iteration. We next update the l-th functional coefficient starting from $l = 1$. Let*

$$\hat{\boldsymbol{\alpha}}_{-l}^{(j)}(U_s) = \left[\hat{\alpha}_1^{(j+1)}(U_s), \cdots, \hat{\alpha}_{l-1}^{(j+1)}(U_s), 0, \hat{\alpha}_{l+1}^{(j)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}^{(j)}(U_s)\right]^{\intercal},$$

$$\hat{\boldsymbol{\alpha}}^{\prime(j)}(U_s) = \left[\hat{\alpha}_1^{\prime(j)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}^{\prime(j)}(U_s)\right]^{\intercal},$$

$$\hat{Y}_{t,-l}^{(j)} = Y_t - \tilde{\mathbf{X}}_t\hat{\boldsymbol{\alpha}}_{-l}^{(j)}(U_s) - \tilde{\mathbf{X}}_t\hat{\boldsymbol{\alpha}}^{\prime(j)}(U_s)(U_t - U_s),$$

$$\tilde{\mathbf{E}}_{nl} = \left(\tilde{E}_{nl,1}, \cdots, \tilde{E}_{nl,n}\right)^{\intercal}, \quad \tilde{E}_{nl,s} = \frac{2}{nh}\sum_{t=1}^{n} \tilde{X}_{t,l|\tilde{K}}\hat{Y}_{t,-l}^{(j)} K_h(U_t, U_s).$$

If $\|\tilde{\mathbf{E}}_{nl}\| < p'_{\lambda_1}\big(\|\tilde{A}_l\|\big)$, *we update* $\hat{A}_l^{(j+1)} = \mathbf{0}$, *otherwise,*

$$\hat{A}_l^{(j+1)} = \Big[\tilde{\mathbf{\Omega}}_{nl}(0) + p'_{\lambda_1}\big(\|\tilde{A}_l\|\big)\mathbf{I}_n/c_l\Big]^{-1}\tilde{\mathbf{E}}_{nl},$$

where $\mathbf{I}_n$ *is an* $n \times n$ *identity matrix,* $c_l = \|\hat{A}_l^{(j)}\|$ *if* $\|\hat{A}_l^{(j)}\| \neq 0$, *and* $c_l = \max_{k\neq l}\|\hat{A}_k^{(j)}\|$ *if* $\|\hat{A}_l^{(j)}\| = 0$.

3. *Update the derivative of the* l*-th functional coefficient starting from* $l = 1$. *Let*

$$\hat{\boldsymbol{\alpha}}^{(j+1)}(U_s) = \Big[\hat{\alpha}_1^{(j+1)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}^{(j+1)}(U_s)\Big]^{\intercal},$$

$$\hat{\boldsymbol{\alpha}}_{-l}^{\prime(j)}(U_s) = \Big[\hat{\alpha}_1^{\prime(j+1)}(U_s), \cdots, \hat{\alpha}_{l-1}^{\prime(j+1)}(U_s), 0, \hat{\alpha}_{l+1}^{\prime(j)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}^{\prime(j)}(U_s)\Big]^{\intercal},$$

$$\check{Y}_{t,-l}^{(j)} = Y_t - \tilde{\mathbf{X}}_t\hat{\boldsymbol{\alpha}}^{(j+1)}(U_s) - \tilde{\mathbf{X}}_t\hat{\boldsymbol{\alpha}}_{-l}^{\prime(j)}(U_s)(U_t - U_s),$$

$$\check{\mathbf{E}}_{nl} = \big(\check{E}_{nl,1}, \cdots, \check{E}_{nl,n}\big)^{\intercal},$$

$$\check{E}_{nl,s} = \frac{2}{nh}\sum_{t=1}^{n}\tilde{X}_{t,l|\tilde{K}}\check{Y}_{t,-l}^{(j)}[(U_t - U_s)/h]K_h(U_t, U_s).$$

If $\|\check{\mathbf{E}}_{nl}\| < p'_{\lambda_2}\big(\|\tilde{D}_l\|\big)$, *we update* $\hat{B}_l^{(j+1)} = \mathbf{0}$, *otherwise,*

$$h\hat{B}_l^{(j+1)} = \Big[\tilde{\mathbf{\Omega}}_{nl}(2) + p'_{\lambda_2}\big(\|\tilde{D}_l\|\big)\mathbf{I}_n/d_l\Big]^{-1}\check{\mathbf{E}}_{nl},$$

where $d_l = \|h\hat{B}_l^{(j)}\|$ *if* $\|\hat{B}_l^{(j)}\| \neq 0$, *and* $d_l = \max_{k\neq l}\|h\hat{B}_k^{(j)}\|$ *if* $\|\hat{B}_l^{(j)}\| = 0$.

4. *Repeat Steps 2 and 3 until the convergence of the estimates.*

Our numerical studies in Section 4.3 and Section 4.4 below show that the above iterative procedure has a reasonably good finite-sample performance.

Chapter 4

Numerical Study

In this section, we conduct two Monte-Carlo simulation examples and two real data analysis to evaluate the finite-sample performance of the proposed method.

4.1 Simulation Example I

Example I. Consider the following functional-coefficient model:

$$Y_t = \sum_{j=1}^{p} \beta_j(U_t) X_{tj} + \sigma\varepsilon_t, \quad t = 1, \cdots, n, \tag{4.1}$$

where the random covariate vector $\mathbf{X}_t = (X_{t1}, \cdots, X_{tp})^\intercal$ with $p = 20$ is independently generated from a multiple normal distribution with zero mean,

unit variance and correlation ϱ being either 0 or 0.25, the univariate index variable U_t is independently generated from a uniform distribution $\mathsf{U}[0,1]$, the random error ε_t is independently generated from the standard normal distribution and $\sigma = 0.5$. The homogeneity structure on model (4.1) is defined as follows:

- $\beta_{4(k-1)+j}(\cdot) = \alpha_k(\cdot)$ for $k = 1, 2$ and $j = 1, 2, 3, 4$,
- $\beta_{4(k+1)+j}(\cdot) \equiv c_k$ for $k = 1, 2, 3$ and $j = 1, 2, 3, 4$,
- $\alpha_1(u) = \sin(2\pi u)$,
- $\alpha_2(u) = (1 + \delta)\sin(2\pi u)$,
- $c_1 = 0.5$,
- $c_2 = 0.5 + \delta$,
- $c_3 = 0$,

and $\delta = 0.4$ or 0.8. The sample size n is 200, 500 or 1000, and the replication number N is 500.

The above homogeneity structure shows that there are five clusters among the functional and constant coefficients in model (4.1) and the size of each cluster is the same. We first use the kernel smoothing method to obtain the preliminary nonparametric estimates of the functional coefficients

$\beta_j(\cdot), j = 1, \cdots, 20$, where the Epanechnikov kernel $K(z) = \frac{3}{4}(1-z^2)_+$ is used and the optimal bandwidth is determined by the cross-validation criterion in Section 3.3. The homogeneity and semi-varying coefficient structure in model (4.1) is ignored in this preliminary nonparametric estimation procedure. A combination of the kernel-based clustering method and the generalised information criterion in Section 3.1 is then used to estimate the latent homogeneity structure in the simulation. In order to evaluate the clustering performance, we consider two commonly-used measurements: Normalised Mutual Information (NMI) and Purity, both of which can be used to examine how close are the estimated set of clusters to the true set of clusters. Letting $\mathcal{C}_1 = \left\{\mathcal{C}_1^1, \cdots, \mathcal{C}_{K_1}^1\right\}$ and $\mathcal{C}_2 = \left\{\mathcal{C}_1^2, \cdots, \mathcal{C}_{K_2}^2\right\}$ be two sets of disjoint clusters of $(1, 2, \cdots, p)$, the NMI measure is defined as

$$\mathsf{NMI}(\mathcal{C}_1, \mathcal{C}_2) = \frac{I(\mathcal{C}_1, \mathcal{C}_2)}{(H(\mathcal{C}_1) + H(\mathcal{C}_2)/2},$$

where $I(\mathcal{C}_1, \mathcal{C}_2)$ is mutual information between $\mathcal{C}_1$ and $\mathcal{C}_2$:

$$I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{k=1}^{K_1} \sum_{j=1}^{K_2} \left(\frac{|\mathcal{C}_k^1 \cap \mathcal{C}_j^2|}{p} \right) \log \left(\frac{p|\mathcal{C}_k^1 \cap \mathcal{C}_j^2|}{|\mathcal{C}_k^1||\mathcal{C}_j^2|} \right),$$

$H(\mathcal{C}_1) = \frac{|\mathcal{C}_1|}{p} \times \log(\frac{|\mathcal{C}_1|}{p})$ and $H(\mathcal{C}_2) == \frac{|\mathcal{C}_2|}{p} \times \log(\frac{|\mathcal{C}_2|}{p})$ are the entropy of $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively. The NMI measure takes a value between 0 and 1 with

a larger value indicating that the two sets of clusters are closer. The Purity measure is defined by

$$\mathsf{Purity}(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{p} \sum_{k=1}^{K_1} \max_{1 \le j \le K_2} |\mathcal{C}_k^1 \cap \mathcal{C}_j^2|. \tag{4.2}$$

It is easy to find that the Purity measure also takes a value between 0 and 1, and $\mathsf{Purity}(\mathcal{C}_1, \mathcal{C}_2) = 1$ means that $\mathcal{C}_1$ is exactly the same as $\mathcal{C}_2$. Table 4.1 below summarises the estimation of cluster number in 500 replications and Table 4.2 below gives the means and standard errors (in parentheses) for the NMI and Purity measurements. From Table 4.1, we can find that the number of clusters in general can be accurately estimated and it improves significantly when the sample size increases from 200 to 500. Table 4.2 shows that if there is no correlation among the random covariates, the NMI and Purity values are close to one even when the sample size is as small as 200. The increase of the correlation ϱ from 0 to 0.25 has an impact on small-sample simulation performance of the proposed clustering approach in particular when the sample size is 200.

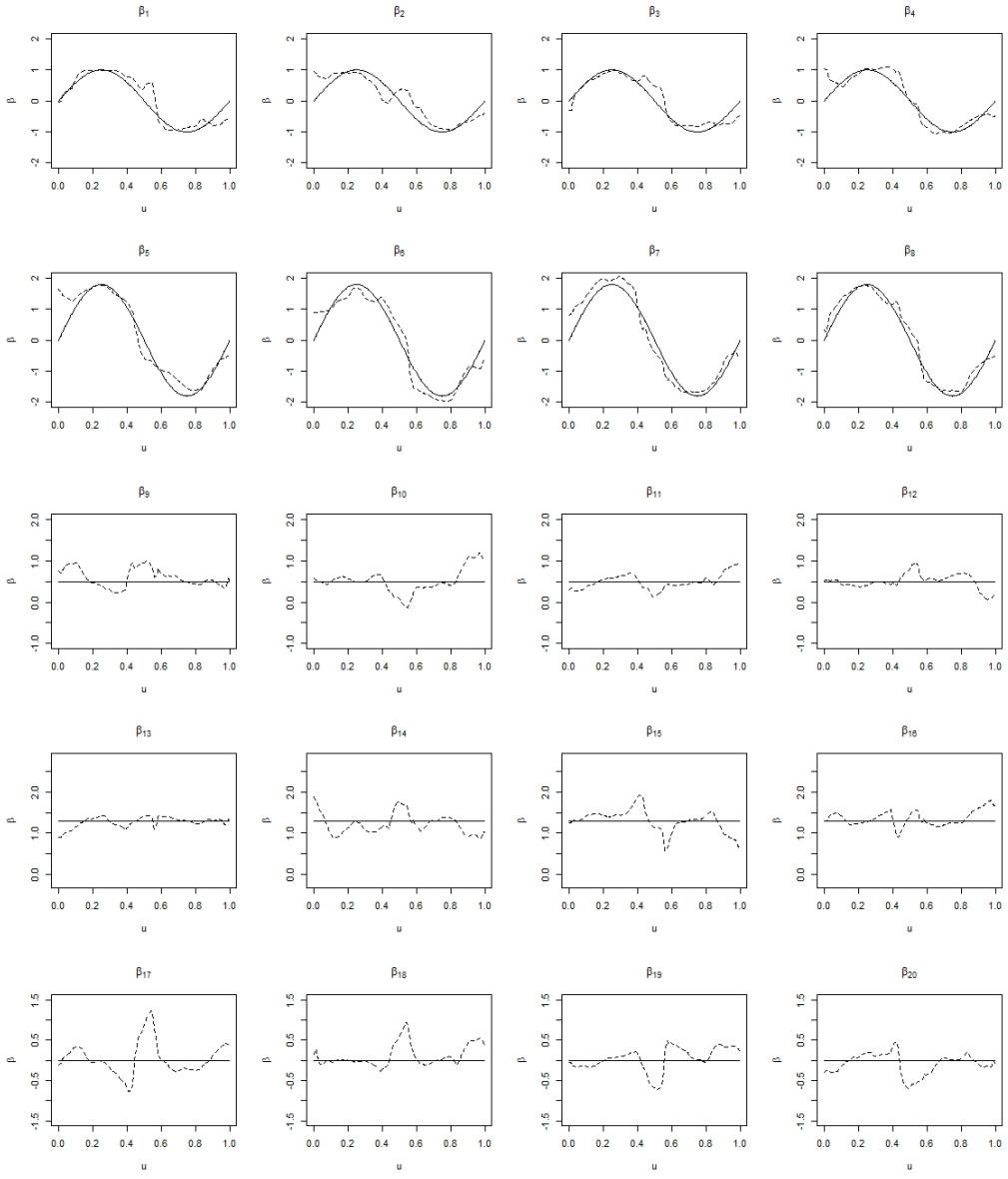

Figure 4.1: The preliminary kernel estimation of Example I (n=200)

The preliminary kernel estimation of the functional coefficients from a typical realisation of model (4.1) with "HS I" when the sample size $n = 200$ and $\delta = 0.8$. The solid lines are true coefficient functions and the dash lines are estimated curves.

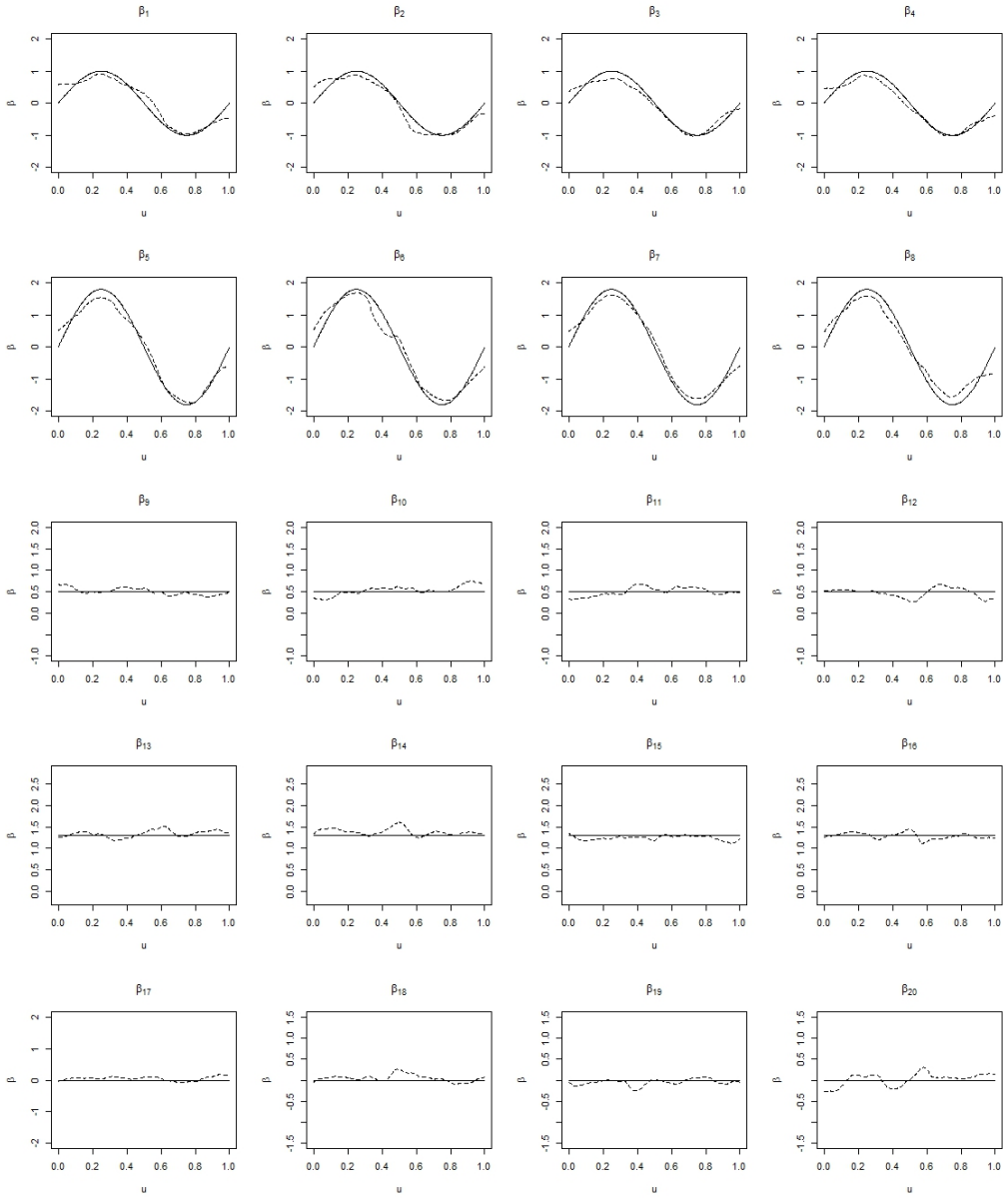

Figure 4.2: The preliminary kernel estimation of Example I (n=500)

The preliminary kernel estimation of the functional coefficients from a typical realisation of model (4.1) with "HS I" when the sample size $n = 500$ and $\delta = 0.8$. The solid lines are true coefficient functions and the dash lines are estimated curves.

Table 4.1: Result on estimation of cluster number in Example I

δ	ϱ	n	$\mathsf{K}=3$	$\mathsf{K}=4$	$\mathsf{K}=5$	$\mathsf{K}=6$
0.4	0	200	0	99	401	0
		500	0	0	500	0
		1000	0	0	500	0
0.4	0.25	200	122	125	253	0
		500	0	0	500	0
		1000	0	0	500	0
0.8	0	200	0	3	497	0
		500	0	0	500	0
		1000	0	0	500	0
0.8	0.25	200	8	73	418	1
		500	0	0	500	0
		1000	0	0	500	0

Table 4.2: Result on the NMI and Purity measurements in Example I

		$\delta=0.4$		$\delta=0.8$	
ϱ	n	NMI	Purity	NMI	Purity
0	200	0.83 (0.18)	0.87 (0.16)	0.94 (0.13)	0.96 (0.12)
	500	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	1000	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
0.25	200	0.46 (0.12)	0.55 (0.09)	0.89 (0.17)	0.90 (0.14)
	500	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	1000	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

This table shows the accuracy of estimated classification structure for different sample. The values in parenthesis report standard error based on 500 replications.

We finally identify the clusters with zero coefficients and non-zero constant coefficients by using the penalised method introduced in Section 3.4. The tuning parameters in the penalty term are chosen by the GIC given in Section 3.3, where for simplicity we let $\lambda_1 = \lambda_2 = \lambda$ (which is reasonable due to Assumption 9). In order to measure the accuracy of the shrinkage

method, we use the Mean Absolute Estimation Error (MAEE) defined by

$$\mathsf{MAEE}(P) = \frac{1}{n\tilde{K}} \sum_{t=1}^{n} \sum_{k=1}^{\tilde{K}} |\hat{\alpha}_{\lambda,k}(U_t) - \alpha_k(U_t)|,$$

where $\alpha_k(\cdot)$, $k = 1, \cdots, \tilde{K}$, are true functional (or constant) coefficients and $\hat{\alpha}_{\lambda,k}(\cdot)$ are their penalised estimates. Similarly, for the preliminary kernel estimation, the corresponding MAEE is defined by

$$\mathsf{MAEE}(K) = \frac{1}{np} \sum_{t=1}^{n} \sum_{j=1}^{p} |\tilde{\beta}_j(U_t) - \beta_j(U_t)|, \quad p = 20,$$

where $\tilde{\beta}_j(\cdot)$, $j = 1, \cdots, 20$, are the preliminary kernel estimates of the true coefficient functions $\beta_j(\cdot)$. Table 4.8 below reports the median of the MAEE values over 500 replications for both the preliminary kernel estimation and the proposed semiparametric shrinkage method. The result in the table shows that, after identifying the homogeneity and semi-varying coefficient structure, the MAEE values of the semiparametric penalised estimation are much smaller than those by directly applying the nonparametric kernel estimation. In addition, both estimation methods improve their performance (with decreasing MAEE values) as the sample size increases, and performance becomes slightly worse when the correlation between the random covariates increases from 0 to 0.25.

Table 4.3: Median of MAEE values over 500 replications in Example I

		$\delta = 0.4$		$\delta = 0.8$	
ϱ	n	MAEE(P)	MAEE(K)	MAEE(P)	MAEE(K)
0	200	0.063	0.171	0.063	0.243
	500	0.029	0.081	0.026	0.091
	1000	0.018	0.060	0.018	0.063
0.25	200	0.064	0.241	0.068	0.269
	500	0.045	0.119	0.023	0.137
	1000	0.023	0.078	0.034	0.087

Apart from above MAEE results, we also apply the out-of-sample predictive performance between the proposed approach and the preliminary kernel estimation. Here we randomly split the full sample into the training set (containing 90 % of observations for model estimation) and the testing set (containing the remaining 10 % observations for evaluating the model predictive capacity). The predictive performance is measured by Mean Squared Prediction Error (MSPE), which is defined by

$$\mathsf{MSPE} = \frac{1}{n_\star} \sum_{i=1}^{n_\star} \left(Y_i^\star - \hat{Y}_i^\star \right)^2, \tag{4.3}$$

where $n_\star = 20, 50, 100$ are the testing sample size, $Y_i^\star$ is the true value of response variable in the testing sample, and $\hat{Y}_i^\star$ is the fitted value of $Y_i^\star$ using the model estimation in the training sample. Table 4.4 below reports the means of the MSPE values over 500 times of random sample splitting, where MSPE(P) denotes the MSPE using the proposed kernel clustering

analysis and penalised estimation method in the training set, and MSPE(K) denotes the MSPE using the preliminary kernel estimation in the training set. From the Table 4.4, the MSPE(P) values are significantly smaller than the MSPE(K) values. This comparison result shows that the simplified functional-coefficient model via the developed kernel-based clustering and structure identification provides more accurate out-of-sample prediction result.

Table 4.4: Median of MSPE over 500 replications in Example I

		$\delta = 0.4$		$\delta = 0.8$	
ϱ	n	MSPE(P)	MSPE(K)	MSPE(P)	MSPE(K)
0	200	0.203	0.325	0.186	0.263
	500	0.142	0.211	0.106	0.201
	1000	0.129	0.170	0.096	0.157
0.25	200	0.264	0.361	0.217	0.339
	500	0.159	0.219	0.129	0.267
	1000	0.123	0.198	0.114	0.187

4.2 Simulation Example II

Example II. We still consider model (4.1) with the following homogeneity structure:

- $\beta_1(\cdot) = \alpha_1(\cdot)$,
- $\beta_j(\cdot) = \alpha_2(\cdot)$ for $j = 2$ and 3,
- $\beta_j(\cdot) \equiv c_1$ for $j = 4, \cdots, 7$,

- $\beta_j(\cdot) \equiv c_2$ for $j = 8, \cdots, 13$,
- $\beta_j(\cdot) \equiv c_3$ for $j = 14, \cdots, 20$.
- $\alpha_1(u) = \sin(2\pi u)$,
- $\alpha_2(u) = (1 + \delta)\sin(2\pi u)$,
- $c_1 = 0.5$,
- $c_2 = 0.5 + \delta$,
- $c_3 = 0$.

The data generating processes for the random covariates $\mathbf{X}_t$, the index variable U_t and the model error ε_t are the same as those in Example 4.1. The definitions of $\alpha_i(\cdot)$ and c_i are also the same as those in the previous example.

Tables 4.5 and 4.6 report the simulation result for the estimated latent homogeneity structure and Table 4.7 and 4.8 report the medians of the MAEE and MSPE values (for both the preliminary kernel estimation and penalised local linear estimation) in 500 replications. Although the size of clusters varies in this example, the simulation results are generally similar to those obtained in Example I. Details are omitted here to save the space.

Table 4.5: Result on estimation of cluster number in Example II

δ	ϱ	n	$\mathsf{K}=3$	$\mathsf{K}=4$	$\mathsf{K}=5$	$\mathsf{K}=6$
0.4	0	200	0	174	326	0
		500	0	0	500	0
		1000	0	0	500	0
0.4	0.25	200	1	402	97	0
		500	0	0	500	0
		1000	0	0	500	0
0.8	0	200	0	3	497	0
		500	0	0	500	0
		1000	0	0	500	0
0.8	0.25	200	0	36	454	10
		500	0	0	500	0
		1000	0	0	500	0

Table 4.6: Result on the NMI and Purity measurements in Example II

		$\delta = 0.4$		$\delta = 0.8$	
ϱ	n	NMI	Purity	NMI	Purity
0	200	0.83 (0.17)	0.84 (0.13)	0.94 (0.13)	0.96 (0.12)
	500	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	1000	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
0.25	200	0.49 (0.12)	0.54 (0.09)	0.97 (0.06)	0.99 (0.03)
	500	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
	1000	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)

Table 4.7: Median of MAEE values over 500 replications in Example II

		$\delta = 0.4$		$\delta = 0.8$	
ϱ	n	MAEE(P)	MAEE(K)	MAEE(P)	MAEE(K)
0	200	0.073	0.163	0.051	0.127
	500	0.029	0.068	0.025	0.075
	1000	0.019	0.053	0.018	0.056
0.25	200	0.055	0.141	0.057	0.159
	500	0.034	0.076	0.026	0.085
	1000	0.021	0.065	0.018	0.056

Table 4.8: Median of MSPE over 500 replications in Example II

		$\delta = 0.4$		$\delta = 0.8$	
ϱ	n	MSPE(P)	MSPE(K)	MSPE(P)	MSPE(K)
0	200	0.212	0.329	0.180	0.257
	500	0.147	0.225	0.112	0.209
	1000	0.133	0.173	0.106	0.164
0.25	200	0.268	0.369	0.225	0.326
	500	0.154	0.218	0.132	0.255
	1000	0.134	0.206	0.117	0.190

4.3 Real Data Analysis I

In this section, we apply the developed model and methodology to two real data sets: the Boston house price data and the plasma beta-carotene level data. These two data sets have been extensively analysed in some existing studies where the functional-coefficient model is usually recommended. However, it is not clear whether certain homogeneity structure among the functional coefficients exists. This motivates us to further examine the modelling structure via the kernel-based clustering method and penalised approach introduced in Chapter 3.

Real data I.

We first apply the developed model and methodology to the well-known Boston house price data. In last two decades, real estate plays a significant role in the world economy, especially in the United States. Boston is the

largest and one of the oldest cities in the US with a population of over 685,000. Loads of world-famous universities and research institutes are located in Boston and surrounding areas. As a supreme financial center, Boston has some of the highest home prices of major cities in the US and it is quite value to do some research on Boston house price. The data set we use has been previously analysed in some existing studies (c.f., Fan and Huang, 2005; Wang and Xia, 2009). The meaning of variables we use as follows:

- *MEDV: the median value of owner-occupied homes in the unit of US$* 1000.
- *CRIM: the crime rate per capita by town.*
- *RM: the average number of rooms per dwelling.*
- *PTRATIO: the ratio of pupil-teacher by town.*
- *TAX: the full-value property-tax rate per US$* 10000.
- *NOX: the nitric oxides concentration per* 10 *million.*
- *AGE: the proportion of owner-occupied units built prior to* 1940.
- *LSTAT: the percentage of lower status of the population.*
- *INT: Intercept.*

As in the literature, we select MEDV (the median value of owner-occupied homes in the unit of US$ 1000) as the response variable. The candidate explanatory variables include CRIM (the crime rate per capita by town), RM (the average number of rooms per dwelling), PTRATIO (the ratio of pupil-teacher by town), TAX (the full-value property-tax rate per US$ 10000), NOX (the nitric oxides concentration per 10 million), AGE (the proportion of owner-occupied units built prior to 1940) and INT (the intercept). The LSTAT (the percentage of lower status of the population) variable is chosen as the index variable U in the functional-coefficient model. The Z-score method is applied to transform the response and explanatory variables (except INT). The LSTAT variable is min-max normalization transformed so that its distribution is $\mathsf{U}(0,1)$, consistent with the assumption made on the asymptotic theory.

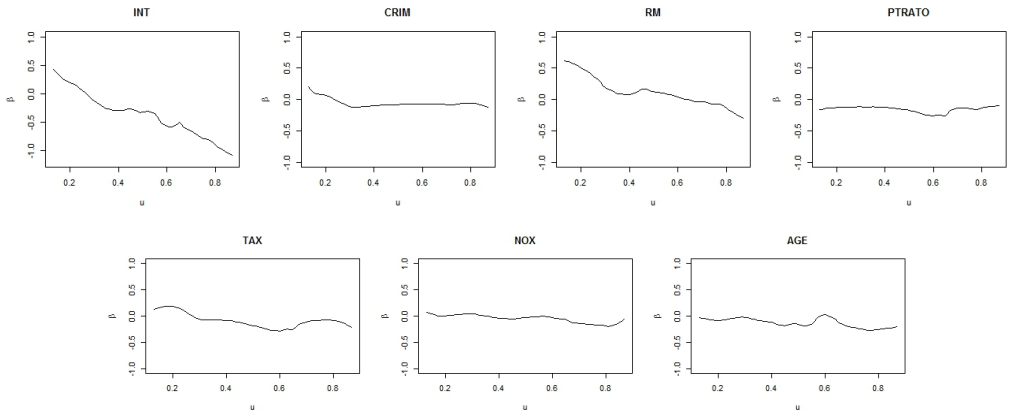

Figure 4.3: The preliminary kernel estimated curves of the functional coefficients.

Figure 4.3 plots the preliminary kernel estimated curves for the functional coefficients corresponding to the intercept and six explanatory variables, where the bandwidth is $h_{opt} = 0.13$ determined by the leave-one-out cross-validation method. From Figure 4.3, we observe that the coefficients for CRIM, PTRATIO, TAX, NOX and AGE have similar functional pattern (close to the horizontal line), indicating that they might take constant values. This is confirmed by using the methodology proposed in Section 3.2. The kernel-based cluster analysis and the generalised information criterion identifies the following three clusters: the functional coefficients corresponding to CRIM, PTRATIO, TAX, NOX and AGE are identical and form one cluster, the functional coefficients corresponding to INT and RM form two other clusters, respectively. Furthermore, the penalised local linear estimation with the optimal tuning parameters chosen as $\lambda_1 = 6.5$ and $\lambda_2 = 3$ suggests that the identical coefficient function for CRIM, PTRATIO, TAX, NOX and AGE has a constant value (-0.023), whereas the coefficient functions for INT and RM changes with the LSTAT variable. The two estimated functional coefficients are given in Figure 4.4. The estimated intercept function is overall decreasing, indicating that the house price would drop as the LSTAT value increases. The estimated functional coefficient associated with the RM variable is mostly positive (in particular when

the LSTAT value is relatively small (relatively high educational status neighborhood), indicating a positive relationship between the house price and RM. By applying the methodology proposed in my thesis, there are only two nonparametric components and one parameter in the final semi-varying coefficient model, which is much simpler than the pure functional-coefficient model considered in some existing literature.

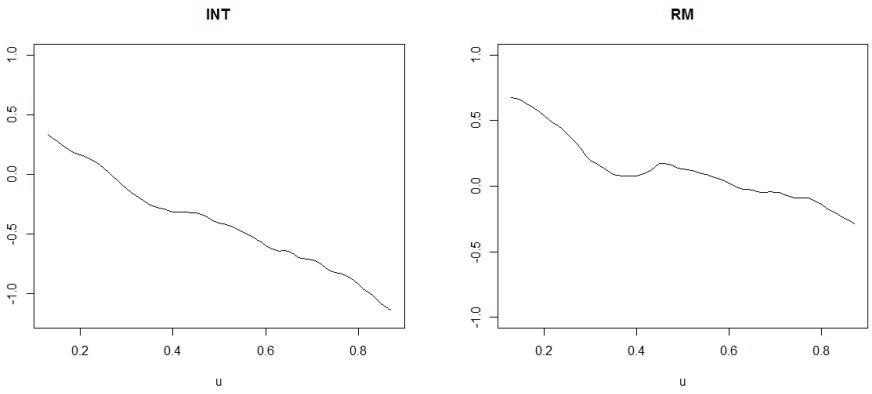

Figure 4.4: The estimated functional coefficients of INT and RM.

We next compare the out-of-sample predictive performance between the proposed approach and the preliminary kernel estimation. We still randomly split the full sample into the training set (containing 456 observations for model estimation) and the testing set (containing the remaining 50 observations for evaluating the model predictive capacity). The predictive performance is measured by Mean Squared Prediction Error (MSPE), which is defined in model (4.3) where the testing sample size $n_\star = 50$, $Y_i^\star$ is the true value of response variable in the testing sample. Table 4.9 below reports

the means of the MSPE values over 200 times of random sample splitting, where MSPE(P) denotes the MSPE using the proposed kernel clustering analysis and penalised estimation method in the training set, and MSPE(K) denotes the MSPE using the preliminary kernel estimation in the training set. We consider the bandwidth values starting from 0.04 to 0.16 (with equal distance 0.02), covering $h_{opt} = 0.08$ which is the optimal bandwidth for the penalised local linear estimation. From the table, we can see that when the bandwidth is close to or smaller than the optimal bandwidth $h_{opt} = 0.08$ (for penalised estimation), the MSPE(P) values are significantly smaller than the MSPE(K) values. Only when $h = 0.12$ (close to $h_{opt} = 0.13$, the optimal bandwidth for the preliminary kernel estimation), the MSPE(K) value is smaller than the MSPE(P) value. This comparison result shows that the simplified functional-coefficient model via the developed kernel-based clustering and structure identification provides more accurate out-of-sample prediction result.

Table 4.9: MSPE values over 200 times of random sample splitting in Real Data I

	$h = 0.04$	$h = 0.06$	$h = 0.08$	$h = 0.10$
MSPE(P)	0.354	0.343	0.324	0.352
MSPE(K)	0.872	0.708	0.575	0.368
	$h = 0.12$	$h = 0.14$	$h = 0.16$	$h = 0.18$
MSPE(P)	0.332	0.334	0.335	0.347
MSPE(K)	0.277	0.673	0.670	0.716

4.4 Real Data Analysis II

Real Data II.

We next apply the developed method to analyse the plasma beta-carotene level data, which have been previously studied by Nierenberg et al.(1989), Wang and Li (2009) and Kai, Li and Zou (2011). Plasma can transport to retinol, which becomes one kind of vitamin A alcohol. Beta-carotene is an antioxidant that converts to vitamin A and plays a crucial role in health. Therefore, it is meaningful to explore the relationship between dietary factors and personal characteristics.

Existing studies have suggested that low dietary intake or low plasma concentrations of retinol, beta-carotene, or other carotenoids might be associated with high risk of developing certain types of cancer. However, quite few studies have explored the determinants of plasma concentrations of these micronutrients. Here, followed by existing literature I have mentioned above, we intend to investigate the relationship between personal characteristics and dietary factors, plasma concentrations of retinol, beta-carotene and other carotenoids. We start the empirical analysis with the functional-coefficient model with given variables,

- *AGE: Age (years).*

- *SMOKSTAT: Smoking status (1 = Never, 2 = Former, 3 = Current Smoker).*
- *QUETELET: Quetelet (weight/(height^2)).*
- *VITUSE: Vitamin Use (1= Yes, fairly often, 2 = Yes, not often, 3 = No).*
- *CALORIES: Number of calories consumed per day.*
- *FAT: Grams of fat consumed per day.*
- *FIBER: Grams of fiber consumed per day.*
- *ALCOHOL: Number of alcoholic drinks consumed per week.*
- *CHOLESTEROL: Cholesterol consumed (mg per day).*
- *BETADIET: Dietary beta-carotene consumed (mcg per day).*
- *RETDIET: Dietary retinol consumed (mcg per day)*
- *BETAPLASMA: Plasma beta-carotene (ng/ml).*
- *RETPLASMA: Plasma Retinol (ng/ml).*

In this analysis, the response variable is chosen as PBCL and the candidate explanatory variables include AGE, SMOKSTAT, VITUSE, QUETELET,

CALORIES, FAT, FIBE, ALCOHO, CHOLESTEROLand INT (the intercept). Followed by Kai and Zou (2011), the Z-score method are used to transform the response and random explanatory variables and the DBC variable is min-max normalisation transformed so that its distribution follows the uniform distribution $\mathsf{U}(0,1)$. The index variable U is chosen as DBC (dietary beta-carotene consumed per day).

In the preliminary kernel estimation, the Epanechnikov kernel $K(z) = \frac{3}{4}(1-z^2)_+$ is used and the optimal bandwidth is determined via the cross-validation method in Chapter 3 with $h_{opt} = 0.18$. The kernel-based clustering method and penalised local linear estimation (with the tuning parameters $\lambda_1 = 4$ and $\lambda_2 = 2$ chosen by the GIC method) are combined to explore the homogeneity structure among the functional coefficients. The following three clusters are identified: the functional coefficient for INT (i.e., the intercept function) forms the first cluster, functional coefficients for SMOKSTAT and QUETELET have the same pattern and form the second cluster, and the functional coefficients for the remaining seven covariates form the third cluster (with coefficients being zero). Figure 4.5 plots the estimated curves for the two significant functional coefficients. The estimated intercept function is overall increasing, indicating that the plasma beta-carotene level increases as the index variable DBC increases. The estimated functional coefficient

associated with the covariates QUETELET and SMOKSTAT are in general negative, indicating a negative relationship between the plasma beta-carotene level and the combined covariates QUETELET and SMOKSTAT.

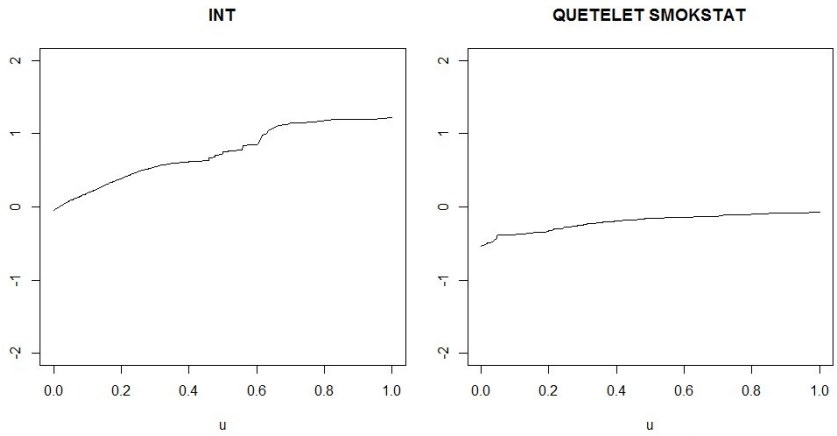

Figure 4.5: The estimated curves of the two significant functional coefficients corresponding to INT and QUETELET+SMOKSTAT, respectively.

Moreover, we further compare the out-of-sample predictive performance between the proposed approach and the preliminary kernel estimation. We randomly divide the full sample (with 273 observations) into the training set (containing 243 observations) and the testing data (containing 30 observation), and repeat such a random sample splitting 200 times to avoid randomness. The MSPE measurement defined in (4.3) is used to evaluate the predictive performance with the result reported in Table 4.10 below. From the table, we find that the proposed kernel-based clustering and semi-parametric shrinkage method usually outperforms the preliminary kernel estimation method (ignoring the latent homogeneity structure) in terms of

predictive measurement.

Table 4.10: MSPE values over 200 times of random sample splitting in Real Data II

	$h = 0.12$	$h = 0.15$	$h = 0.18$	$h = 0.21$	$h = 0.24$
MSPE(P)	1.207	1.048	1.017	1.027	1.024
MSPE(K)	2.214	2.029	1.583	1.189	0.963

Chapter 5

Related Asymptotic Theorems

5.1 Asymptotic Theorems

In this section, we give the asymptotic theorems for the proposed clustering and semiparametric penalised methods. We start with some regularity conditions, some of which might be weakened at the expense of more lengthy proofs.

Assumption 1. *The kernel function $K(\cdot)$ is a Lipschitz continuous and symmetric probability density function with a compact support $[-1, 1]$.*

Assumption 2(i). *The density function of the index variable U_t, $f_U(\cdot)$, has continuous second-order derivative and is bounded away from zero and infinity on the support.*

(ii). *The functional coefficients* $\boldsymbol{\beta}_0(\cdot)$ *and* $\boldsymbol{\alpha}_0(\cdot) = \left[\alpha_1^0(\cdot), \cdots, \alpha_{K_0}^0(\cdot)\right]^\intercal$ *have continuous second-order derivatives.*

Assumption 3(i). *The* $p \times p$ *matrix* $\boldsymbol{\Sigma}(u) := \mathsf{E}\left(\mathbf{X}_t\mathbf{X}_t^\intercal | U_t = u\right)$ *is twice continuously differentiable and positive definite for any* $u \in [0,1]$. *Furthermore,*

$$0 < \inf_{u\in[0,1]} \lambda_{\min}(\boldsymbol{\Sigma}(u)) \leq \sup_{u\in[0,1]} \lambda_{\max}(\boldsymbol{\Sigma}(u)) < \infty,$$

where $\lambda_{\min}(\cdot)$ *and* $\lambda_{\max}(\cdot)$ *denote the smallest and largest eigenvalues, respectively.*

(ii). *Let* $(U_t, \mathbf{X}_t, \varepsilon_t)$, $t = 1, \cdots, n$, *be i.i.d. Furthermore, the error* ε_t *is independent of* $(U_t, \mathbf{X}_t)$, $\mathsf{E}[\varepsilon_t] = 0$ *and* $0 < \sigma^2 = \mathsf{E}[\varepsilon_t^2] < \infty$, *and there exists* $0 < \iota_1 < \infty$ *such that* $\mathsf{E}\left(|\varepsilon_t|^{2+\iota_1}\right) + \max_{1\leq i\leq p} \mathsf{E}\left(|X_{ti}|^{2(2+\iota_1)}\right) < \infty$.

Assumption 4(i). *Let the bandwidth* h *and the dimension* p *satisfy*

$$p(\epsilon_n + h^2) = o(1), \quad n^{2\iota_2 - 1}h \to \infty,$$

where $\epsilon_n = \sqrt{\log h^{-1}/(nh)}$ *and* $\iota_2 < 1 - 1/(2 + \iota_1)$.

(ii). *Let*

$$p^{1/2}\left(\epsilon_n + h^2\right) = o(\delta_n), \quad n^{1/2}\delta_n/(\log n)^{1/2} \to \infty,$$

where

$$\delta_n = \min_{1\le k_1\neq k_2\le K_0} \delta_{k_1k_2}, \quad \delta_{k_1k_2} = \int_{\mathcal{U}_h} \left|\alpha^0_{k_1}(u) - \alpha^0_{k_2}(u)\right| f_U(u)du.$$

Remark 1. Assumptions 1–3 are some commonly-used conditions on the kernel estimation of the functional-coefficient models. The strong moment condition on ε_t and $\mathbf{X}_t$ in Assumption 3(ii) is required when applying the uniform asymptotics of some kernel-based quantities. Assumption 4(i) restricts the divergence rate of the regressor dimension and the convergence rate of the bandwidth. In particular, if ι_1 is sufficiently large (i.e., the moment conditions in Assumption 3(ii) becomes stronger), the condition $n^{2\iota_2-1}h \to \infty$ could be close to the conventional condition $nh \to \infty$. Assumption 4(ii) indicates that the difference between two functional coefficients (in different clusters) can be convergent to zero with certain polynomial rate. In particular, when p is fixed, $h = c_h n^{-1/5}$ with $0 < c_h < \infty$, and $\delta_n = n^{-\delta_0}$ with $0 \le \delta_0 < 2/5$, Assumption 4(ii) would be automatically satisfied.

Theorem 1. *Suppose that Assumptions 1–4 are satisfied and K_0 is known*

a priori. Then we have

$$\mathsf{P}\left(\left\{\tilde{\mathcal{C}}_k,\ k=1,\cdots,K_0\right\} \neq \left\{\mathcal{C}_k^0,\ k=1,\cdots,K_0\right\}\right) = o(1) \tag{5.1}$$

when the sample size n is sufficiently large, where $\tilde{\mathcal{C}}_k$ is defined in Section 3.1 and $\mathcal{C}_k^0$ is defined in (1.2).

Remark 2. The above theorem shows the consistency of the agglomerative hierarchical clustering method proposed in Section 3.1 when the number of clusters is known a priori, i.e., with probability approaching one, the K_0 clusters can be correctly specified. It is similar to Theorem 3.1 in (Vogt and Linton 2017) which gives the consistency of classification of nonparametric univariate functions in the longitudinal data setting by using the nonparametric segmentation method.

We next derive the consistency for the information criterion on estimating the number of clusters which is usually unknown in practice. Some further notation and assumptions are needed. Define

$$\mathbf{X}_{t,K_0} = \left(X_{t,1|K_0},\cdots,X_{t,K_0|K_0}\right)^\intercal \quad \text{with} \quad X_{t,k|K_0} = \sum_{j\in\mathcal{C}_k^0} X_{tj},$$

and

$$\mathbf{\Sigma}_{X|K_0}(u) = \mathsf{E}\left[\mathbf{X}_{t,K_0}\mathbf{X}_{t,K_0}^\intercal | U_t = u\right], \quad u\in[0,1].$$

Similarly, we can define $\mathbf{\Sigma}_{X|K}(u)$ when $K > K_0$ and there are further splits on at least one of $\mathcal{C}_k^0$, $k = 1, \cdots, K_0$. Define the event:

$$\mathbf{C}_n(K_0) = \left\{ \left[\tilde{\mathcal{C}}_k,\ k = 1, \cdots, K_0\right] = \left[\mathcal{C}_k^0,\ k = 1, \cdots, K_0\right] \right\}. \tag{5.2}$$

From (5.1) in Theorem 1, we have $\mathsf{P}\left(\mathbf{C}_n(K_0)\right) \to 1$ as $n \to \infty$. Conditional on the event $\mathbf{C}_n(K_0)$, when the number of clusters K is smaller than K_0, two or more clusters of $\mathcal{C}_k^0$, $k = 1, \cdots, K_0$, are falsely merged, which results in K clusters denoted by $\mathcal{C}_{1|K}, \cdots, \mathcal{C}_{K|K}$, respectively, $1 \leq K \leq K_0 - 1$. With such a clustering result, the functional coefficients in model (1.2) and (1.3) cannot be consistently estimated by the kernel smoothing method, as the model is misspecified. However, we may define the "*quasi*" functional coefficients by

$$\boldsymbol{\alpha}_K(u) = \left[\alpha_{1|K}(u), \cdots, \alpha_{K|K}(u)\right]^{\intercal} = \left[\mathbf{\Sigma}_{X|K}(u)\right]^{-1} \mathbf{\Sigma}_{XY|K}(u), \tag{5.3}$$

where $1 \leq K \leq K_0 - 1$,

$$\mathbf{\Sigma}_{X|K}(u) = \mathsf{E}\left[\mathbf{X}_{t,K}\mathbf{X}_{t,K}^{\intercal} | U_t = u\right], \quad \mathbf{\Sigma}_{XY|K}(u) = \mathsf{E}\left[\mathbf{X}_{t,K} Y_t | U_t = u\right], \tag{5.4}$$

and

$$\mathbf{X}_{t,K} = \left(X_{t,1|K}, \cdots, X_{t,K|K}\right)^{\intercal} \quad \text{with} \quad X_{t,k|K} = \sum_{j \in \mathcal{C}_{k|K}} X_{tj}. \tag{5.5}$$

When $K = K_0$, it is easy to find that the quasi functional coefficients becomes the "*genuine*" functional coefficients conditional on the event $\mathbf{C}_n(K_0)$. Define $\varepsilon_{t,K} = Y_t - \mathbf{X}_{t,K}^{\intercal}\boldsymbol{\alpha}_K(U_t)$ and $\boldsymbol{\varepsilon}_{t1,K} = \mathbf{X}_{t,K}\varepsilon_{t,K}$. By (5.3), it is easy to show that

$$\mathsf{E}\left[\boldsymbol{\varepsilon}_{t1,K}|U_t\right] = \mathbf{0} \quad a.s., \tag{5.6}$$

where $\mathbf{0}$ is a null vector whose dimension might change from line to line. A natural nonparametric estimate of $\boldsymbol{\alpha}_K(\cdot)$ would be $\tilde{\boldsymbol{\alpha}}_K(\cdot)$ defined in (3.4) of Section 3.2, where the order of elements in the latter may have to be re-arranged if necessary. The fact of (5.6) and some smoothness condition on $\boldsymbol{\alpha}(\cdot|K)$ may ensure the uniform consistency of the quasi kernel estimation (see the proof of Theorem 2 in Section 5.2).

Let $\mathcal{A}(K_0)$ be the set of K_0-dimensional twice continuously differentiable functions $\boldsymbol{\alpha}(u) = [\alpha_1(u), \cdots, \alpha_{K_0}(u)]^{\intercal}$ such that at least two elements of $\boldsymbol{\alpha}(u)$ are the identical functions over $u \in [0, 1]$. The following additional assumptions are needed when proving the consistency of the information criterion proposed in Section 3.2.

Assumption 5. *There exists a positive constant c_α such that*

$$\inf_{\boldsymbol{\alpha}(\cdot)\in\mathcal{A}(K_0)} \int_0^1 [\boldsymbol{\alpha}_0(u) - \boldsymbol{\alpha}(u)]^\intercal \boldsymbol{\Sigma}_{X|K_0}(u) [\boldsymbol{\alpha}_0(u) - \boldsymbol{\alpha}(u)] f_U(u) du > c_\alpha. \tag{5.7}$$

Assumption 6 (i). *For any $1 \le K \le \bar{K}$, the $K \times K$ matrix $\boldsymbol{\Sigma}_{X|K}(u)$ is positive definite for $u \in [0,1]$.*

(ii). *For any $1 \le K \le K_0$, the quasi functional coefficient $\boldsymbol{\alpha}_K(\cdot)$ has continuous second-order derivatives.*

Assumption 7. *The bandwidth h and the dimension p satisfy $ph^2 = O(\epsilon_n)$, $nh^6 = o(1)$ and $p = o\left(\min\left\{\epsilon_n^{(\rho-1)/2}, \epsilon_n^{-1/3}\right\}\right)$, where ρ is defined in (3.5).*

Remark 3. Assumptions 5 and 6 are mainly used when deriving the asymptotic lower bound of $\tilde{\sigma}_n^2(K)$ which is involved in the definition of $\mathsf{IC}(K)$ when K is smaller than K_0. The restriction (5.7) in Assumption 5 indicates that the K_0 functional elements in $\boldsymbol{\alpha}_0(\cdot)$ needs to be "sufficiently" distinct. We may show that (5.7) is satisfied if $\inf_{1\le K\le K_0} \inf_{u\in[0,1]} \lambda_{\min}\left(\boldsymbol{\Sigma}_{X|K}(u)\right) > c_1 > 0$ and the Lebesgue measure of $\left\{u \in \mathcal{U} : |\alpha_{k_1}^0(u) - \alpha_{k_2}^0(u)| > c_2 > 0\right\}$ is positive for any $k_1 \ne k_2$. Assumption 6 is required to prove the uniform consistency of the kernel estimation for the quasi functional coefficients. Assumption 7

gives some further restriction on h and p, and indicates that the dimension of the covariates can diverge to infinity at a slow polynomial rate of the sample size n. Theorem 2 below shows that the estimated number of clusters which minimises the IC objective function defined in (3.5) is consistent.

Theorem 2. *Suppose that Assumptions 1–7 are satisfied. Then we have*

$$\mathsf{P}\left(\tilde{K} = K_0\right) \to 1, \tag{5.8}$$

where $\tilde{K}$ is defined in (3.6).

Define

$$A_k^0 = \left[\alpha_k^0(U_1), \cdots, \alpha_k^0(U_n)\right]^\intercal, \quad B_k^0 = \left[\alpha_k^{0\prime}(U_1), \cdots, \alpha_k^{0\prime}(U_n)\right]^\intercal,$$

$$\widehat{A}_k = \left[\widehat{\alpha}_k(U_1), \cdots, \widehat{\alpha}_k(U_n)\right]^\intercal, \quad \widehat{B}_k = \left[\widehat{\alpha}_k'(U_1), \cdots, \widehat{\alpha}_k'(U_n)\right]^\intercal.$$

Without loss of generality, conditional on $\mathbf{C}_n(K_0)$ and $\tilde{K} = K_0$, we assume that $\tilde{\mathcal{C}}_1 = \mathcal{C}_1^0, \cdots, \tilde{\mathcal{C}}_{K_0} = \mathcal{C}_{K_0}^0$, otherwise we only need to re-arrange the order of the elements in $\boldsymbol{\alpha}_0(\cdot) = \left[\alpha_1^0(\cdot), \cdots, \alpha_{K_0}^0(\cdot)\right]^\intercal$ in the relevant asymptotic theorems. For notational simplicity, we also assume that $\alpha_{K_0}^0(\cdot) \equiv 0$ and $\alpha_k^0(\cdot) \equiv \alpha_k^0$ for $k = K_*, \cdots, K_0 - 1$ with $1 < K_* < K_0$, where α_k^0 are non-zero constants (the non-zero constant coefficient does not exist when $K_* = K_0$ and all of the functional coefficients would be constants when $K_* = 1$). For

simplicity, we next assume that all the observations of the index variable U_t, $t = 1, \cdots, n$, are in the set of $\mathcal{U}_h$, to avoid the boundary effect of the kernel estimation, but it can be removed if an appropriate truncation technique such as those in Section 3.1 and Section 3.2 is applied to the penalised local linear estimation. Some additional conditions are needed to derive the sparsity result for the penalised estimation in Section 3.2.

Assumption 8. *For any $k = 1, \cdots, K_0 - 1$, there exists a positive constant c_A such that $\|A_k^0\| \geq c_A\sqrt{n}$ with probability approaching one. When $k = 1, \cdots, K_* - 1$ (with $K_* \geq 2$), there exists a positive constant c_D such that $D_k^0 \geq c_D\sqrt{n}$ with probability approaching one.*

Assumption 9. *Let $p^2nh^5 = O(1)$, and the tuning parameter λ_1 satisfy*

$$\lambda_1 = o(n^{1/2}), \quad n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n + p^4h^{-1/2} = o(\lambda_1). \tag{5.9}$$

The condition (3.9) is also satisfied when λ_1 is replaced by λ_2.

Remark 4. Assumption 8 is a key condition to prove that $\|\tilde{A}_k\|/\sqrt{n}$ and $\tilde{D}_k/\sqrt{n}$ are bounded away from zero with probability approaching one, which together with the definition of the SCAD derivative and $\lambda_1 + \lambda_2 = o(n^{1/2})$ in Assumption 9, indicates that when the functional coefficients or their deviations are significant, the influence of the penalty term in (3.7) can be

asymptotically ignored. For the case when p is fixed and $h = c_h n^{-1/5}$ as discussed in Remark 1, if we choose $\lambda_1 = \lambda_2 = n^{\delta_*}$ with $0.1 < \delta_* < 0.5$, (5.9) in Assumption 9 would be satisfied.

Theorem 3. *Suppose that Assumptions 1–9 are satisfied. Then we have*

$$\mathsf{P}\left(\|\widehat{A}_{K_0}\| = 0,\ \|\widehat{B}_k\| = 0, k = K_*, \cdots, K_0\right) \to 1. \tag{5.10}$$

The above sparsity result for the penalised local linear estimation shows that the zero coefficient and non-zero constant coefficients in the model can be identified asymptotically.

5.2 Proof of Theorems

In this section, I will give the detailed proofs of the main asymptotic results.

Proof of Theorem 1. From the definition of Δ^0_{ij}, we have $\Delta^0_{ij} = 0$ if $i, j \in \mathcal{C}^0_k$; and $\Delta^0_{ij} = \delta_{k_1 k_2}$ if $i \in \mathcal{C}^0_{k_1}$ and $j \in \mathcal{C}^0_{k_2}$ with $1 \le k_1 \ne k_2 \le K_0$, where $\delta_{k_1 k_2}$ is defined in Assumption 4(ii). Note that the true number of clusters, K_0, is assumed to be known in this theorem. Therefore, from the algorithm for the clustering method, to prove (5.1), we only need to prove that

$$\max_{1 \le i,j \le p} \left|\tilde{\Delta}_{ij} - \Delta^0_{ij}\right| = o_P(\delta_n), \quad \delta_n = \min_{1 \le k_1 \ne k_2 \le K_0} \delta_{k_1 k_2}. \tag{5.11}$$

From the definitions of $\tilde{\Delta}_{ij}$ and Δ^0_{ij} in Section 3.1, it is sufficient to show

$$\max_{1\le i\le p}\sup_{u\in\mathcal{U}_h}\left|\tilde{\beta}_i(u)-\beta_i^0(u)\right| = o_P(\delta_n). \tag{5.12}$$

In fact, if (5.12) holds, by the definition of $\tilde{\Delta}_{ij}$ and letting

$$\Delta_{ij} = \frac{1}{n}\sum_{t=1}^{n}\left|\beta_i^0(U_t)-\beta_j^0(U_t)\right|\mathsf{I}(U_t\in\mathcal{U}_h),$$

we have

$$\max_{1\le i,j\le p}\left|\tilde{\Delta}_{ij}-\Delta_{ij}\right| \le 2\max_{1\le i\le p}\sup_{u\in\mathcal{U}_h}\left|\tilde{\beta}_i(u)-\beta_i^0(u)\right| = o_P(\delta_n). \tag{5.13}$$

For the case of $i,j\in\mathcal{C}_k^0$, we readily have $\Delta^0_{ij}=\Delta_{ij}=0$, and thus (5.13) leads to (5.11). On the other hand, uniformly for $i\in\mathcal{C}^0_{k_1}$ and $j\in\mathcal{C}^0_{k_2}$ with $1\le k_1\ne k_2\le K_0$, as $n^{1/2}\delta_n/(\log n)^{1/2}\to\infty$ in Assumption 4(ii), we have

$$|\Delta_{ij}-\delta_{k_1k_2}| = O_P\left(\sqrt{\log n/n}\right) = o_P(\delta_n), \tag{5.14}$$

which together with (5.13), implies that (5.11) holds.

We next prove (5.12). By (1.2) and (3.1), we have

$$\begin{aligned}\tilde{\boldsymbol{\beta}}(u)-\boldsymbol{\beta}_0(u) &= \left[\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}K_h(U_t,u)\right]^{-1}\left[\sum_{t=1}^{n}\mathbf{X}_t\varepsilon_t K_h(U_t,u)\right]+\\ &\quad\left[\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}K_h(U_t,u)\right]^{-1}\left[\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}\boldsymbol{\beta}_t(u)K_h(U_t,u)\right] \qquad (5.15)\end{aligned}$$

where $\boldsymbol{\beta}_t(u)=\boldsymbol{\beta}_0(U_t)-\boldsymbol{\beta}_0(u)$. Let

$$\boldsymbol{\Omega}_n(u)=\frac{1}{nh}\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}K_h(U_t,u),\quad \boldsymbol{\Omega}_0(u)=f_U(u)\mathsf{E}\left[\mathbf{X}_t\mathbf{X}_t^{\intercal}|U_t=u\right],$$

and let $\omega_{n,ij}(u)$ and $\omega_{ij}^0(u)$ be the (i,j)-entry of $\boldsymbol{\Omega}_n(u)$ and $\boldsymbol{\Omega}_0(u)$, respectively. By Assumptions 1, 2(i), 3 and 4(i), and using the uniform consistency results for nonparametric kernel-based estimation such as Theorem B in (Mack and Silverman 1982), we have

$$\max_{1\le i,j\le p}\sup_{u\in\mathcal{U}_h}\left|\omega_{n,ij}(u)-\omega_{ij}^0(u)\right|=O_P\left(h^2+\epsilon_n\right), \qquad (5.16)$$

where $\epsilon_n=\sqrt{\log h^{-1}/(nh)}$. Then, by (5.16) and Assumption 4(ii), we may show that

$$\sup_{u\in\mathcal{U}_h}\left\|\boldsymbol{\Omega}_n(u)-\boldsymbol{\Omega}_0(u)\right\|_F=O_P\left(p(\epsilon_n+h^2)\right)=o_P(1), \qquad (5.17)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Using (5.17), Assumption 3(i) and Weyl's inequality, the smallest eigenvalue of $\mathbf{\Omega}_n(u)$ is positive and bounded away from zero uniformly for $u \in \mathcal{U}_h$, i.e.,

$$\inf_{u\in\mathcal{U}_h} \lambda_{\min}(\mathbf{\Omega}_n(u)) > \zeta_0, \tag{5.18}$$

where ζ_0 is a positive constant.

On the other hand, using the uniform consistency result again, we have

$$\sup_{u\in\mathcal{U}_h} \left\| \frac{1}{nh} \sum_{t=1}^{n} \mathbf{X}_t \varepsilon_t K_h(U_t, u) \right\| = O_P\left(p^{1/2}\epsilon_n\right). \tag{5.19}$$

By Assumption 2(ii), applying Taylor's expansion on $\boldsymbol{\beta}^0(\cdot)$ and noting that the largest eigenvalue of $\mathbf{\Sigma}(u) = \mathsf{E}\left(\mathbf{X}_t\mathbf{X}_t^{\intercal}|U_t = u\right)$ is bounded uniformly for $u \in [0,1]$, we also have

$$\sup_{u\in\mathcal{U}_h} \left\| \frac{1}{nh} \sum_{t=1}^{n} \mathbf{X}_t\mathbf{X}_t^{\intercal} \boldsymbol{\beta}_t(u) K_h(U_t, u) \right\| = O_P\left(p^{1/2}h^2\right). \tag{5.20}$$

Combining (5.15) and (5.18)–(5.20), we have

$$\sup_{u\in\mathcal{U}_h} \left\| \tilde{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_0(u) \right\| = O_P\left(p^{1/2}\epsilon_n + p^{1/2}h^2\right) = o_P(\delta_n), \tag{5.21}$$

which leads to (5.12). Therefore, the proof of Theorem 1 has been completed.

□

Proof of Theorem 2. Recall that

$$\mathbf{C}_n(K_0) = \left\{ \left[\tilde{\mathcal{C}}_k,\ k = 1, \cdots, K_0 \right] = \left[\mathcal{C}_k^0,\ k = 1, \cdots, K_0 \right] \right\},$$

and let $\mathbf{C}_n^c(K_0)$ be the complement of $\mathbf{C}_n(K_0)$. From Theorem 1, we readily have

$$\begin{aligned} \mathsf{P}\left(\tilde{K} = K_0\right) &= \mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n(K_0)\right) + \mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n^c(K_0)\right) \\ &= \mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n(K_0)\right) + o(1). \end{aligned} \tag{5.22}$$

Noting that

$$\mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n(K_0)\right) =$$

$$\mathsf{P}\left(\tilde{K} = K_0 | \mathbf{C}_n(K_0)\right) \mathsf{P}\left(\mathbf{C}_n(K_0)\right) = \mathsf{P}\left(\tilde{K} = K_0 | \mathbf{C}_n(K_0)\right)(1 + o(1)),$$

to prove (3.8), we only need to show that

$$\mathsf{P}\left(\tilde{K} = K_0 | \mathbf{C}_n(K_0)\right) \to 1. \tag{5.23}$$

From the definition of $\tilde{K}$, (5.23) can be proved if the following result hold:

$$\mathsf{P}\left(\mathsf{IC}(K) > \mathsf{IC}(K_0),\ 1 \le K \ne K_0 \le \bar{K} \mid \mathbf{C}_n(K_0)\right) \to 1. \tag{5.24}$$

We consider (5.24) separately for the two cases: $1 \le K \le K_0 - 1$ and $K_0 + 1 \le K \le \bar{K}$. If $K_0 = 1$, the first case can be ignored. In fact, (5.25) in Proposition 1 below indicates that when $K_0 \le K \le \bar{K}$ and n is sufficiently large, $\mathsf{IC}(K)$ is a strictly increasing function of K, which proves (5.24) for the second case. On the other hand, for $1 \le K \le K_0 - 1$, Proposition 2 shows that $\mathsf{IC}(K) > \log(\sigma^2 + c_\alpha) + o_P(1) > \log(\sigma^2) + o_P(1) = \mathsf{IC}(K_0) + o_P(1)$, which proves (5.24) for the first case. The proof of Theorem 2 has been completed. □

Proposition 1. *Suppose that the conditions of Theorem 2 are satisfied. For $K_0 \le K \le \bar{K}$, conditional on $\mathbf{C}_n(K_0)$, when n is sufficiently large,*

$$\mathsf{IC}(K) = \log(\sigma^2) + K \cdot \left[\frac{\log(nh)}{nh}\right]^{\rho} (1 + o_P(1)), \tag{5.25}$$

where $\sigma^2 = \mathsf{E}[\varepsilon_t^2]$.

Proof. When $K_0 + 1 \le K \le \bar{K}$, conditional on $\mathbf{C}_n(K_0)$, the misclassification issue would not occur although some of $\mathcal{C}_k^0$, $k = 1, \cdots, K_0$, are further split into smaller clusters. Hence, the kernel estimation of the functional

coefficients may be still uniformly consistent, which is to be proved soon. As in Chapter 3, without loss of generality, conditional on $\mathbf{C}_n(K_0)$, we assume that $\tilde{\mathcal{C}}_1 = \mathcal{C}_1^0, \cdots, \tilde{\mathcal{C}}_{K_0} = \mathcal{C}_{K_0}^0$; otherwise we only need to arrange the order of the true functional coefficients. For simplicity of exposition, we next only consider the case of $K = K_0 + 1$ (other cases can be dealt with similarly), and, without loss of generality, further assume that $\mathcal{C}_{K_0}^0$ is split into $\mathcal{C}_{K_0}^*$ and $\mathcal{C}_{K_0+1}^*$, and let

$$\mathbf{X}_{t,K_0+1} = \left(\sum_{j \in \mathcal{C}_1^0} X_{tj}, \cdots, \sum_{j \in \mathcal{C}_{K_0}^*} X_{tj}, \sum_{j \in \mathcal{C}_{K_0+1}^*} X_{tj} \right)^\intercal .$$

Define

$$\boldsymbol{\alpha}_{K_0+1}^*(\cdot) = \left[\alpha_1^0(\cdot), \cdots, \alpha_{K_0}^0(\cdot), \alpha_{K_0}^0(\cdot) \right]^\intercal ,$$

whose corresponding kernel estimation is defined by

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{K_0+1}(u_0) &= \left[\sum_{t=1}^n \mathbf{X}_{t,K_0+1} \mathbf{X}_{t,K_0+1}^\intercal K_h(U_t, u_0) \right]^{-1} \left[\sum_{t=1}^n \mathbf{X}_{t,K_0+1} Y_t K_h(U_t, u_0) \right] \\
&= \left[\frac{1}{nh} \sum_{t=1}^n \mathbf{X}_{t,K_0+1} \mathbf{X}_{t,K_0+1}^\intercal K_h(U_t, u_0) \right]^{-1} \left[\frac{1}{nh} \sum_{t=1}^n \mathbf{X}_{t,K_0+1} \varepsilon_t K_h(U_t, u_0) \right] + \\
&\quad \left[\frac{1}{nh} \sum_{t=1}^n \mathbf{X}_{t,K_0+1} \mathbf{X}_{t,K_0+1}^\intercal K_h(U_t, u_0) \right]^{-1} \times \\
&\quad \left[\frac{1}{nh} \sum_{t=1}^n \mathbf{X}_{t,K_0+1} \mathbf{X}_{t,K_0+1}^\intercal \boldsymbol{\alpha}_{K_0+1}^*(U_t) K_h(U_t, u_0) \right] \\
&=: \boldsymbol{\Omega}_{n,K_0+1}^{-1}(u_0) \boldsymbol{\Lambda}_{n,K_0+1,\varepsilon}(u_0) + \boldsymbol{\Omega}_{n,K_0+1}^{-1}(u_0) \boldsymbol{\Lambda}_{n,K_0+1,\alpha}(u_0). \qquad (5.26)
\end{aligned}$$

Following the arguments relevant to the kernel uniform consistency in the proof of Theorem 1 above, we may show that the smallest eigenvalue of $\mathbf{\Omega}_{n,K_0+1}(u)$ is positive uniformly for $u \in [0,1]$ by Assumption 6(i), and furthermore,

$$\sup_{u\in\mathcal{U}_h} \|\mathbf{\Lambda}_{n,K_0+1,\varepsilon}(u)\| = O_P\left(p\epsilon_n\right), \ \sup_{u\in\mathcal{U}_h} \left\|\mathbf{\Lambda}_{n,K_0+1,\alpha}(u) - \boldsymbol{\alpha}^*_{K_0+1}(u)\right\| = O_P\left(p^2h^2\right), \tag{5.27}$$

where ϵ_n is defined in the proof of Theorem 1. Then, by (5.26) and (5.27), we have

$$\sup_{u\in\mathcal{U}_h} \left\|\tilde{\boldsymbol{\alpha}}_{K_0+1}(u) - \boldsymbol{\alpha}^*_{K_0+1}(u)\right\| = O_P\left(p\epsilon_n + p^2h^2\right) = O_P\left(p\epsilon_n\right) \tag{5.28}$$

as $ph^2 = O(\epsilon_n)$ in Assumption 7.

Letting $I_t = \mathsf{I}(U_t \in \mathcal{U}_h)$, conditional on $\mathbf{C}_n(K_0)$ and that $\mathcal{C}^0_{K_0}$ is split into $\mathcal{C}^*_{K_0}$ and $\mathcal{C}^*_{K_0+1}$, we have

$$\begin{aligned}
\tilde{\sigma}^2_n(K_0+1) &= \frac{1}{n_h}\sum_{t=1}^{n}\left[Y_t - \mathbf{X}^{\intercal}_{t,K_0+1}\tilde{\boldsymbol{\alpha}}_{K_0+1}(U_t)\right]^2 I_t \\
&= \frac{1}{n_h}\sum_{t=1}^{n}\left[\varepsilon_t - \mathbf{X}^{\intercal}_{t,K_0+1}\left(\tilde{\boldsymbol{\alpha}}_{K_0+1}(U_t) - \boldsymbol{\alpha}^*_{K_0+1}(U_t)\right)\right]^2 I_t \\
&= \frac{1}{n_h}\sum_{t=1}^{n}\varepsilon^2_t I_t + \frac{1}{n_h}\sum_{t=1}^{n}\varpi^2_t(K_0+1)I_t - \\
&\quad \frac{2}{n_h}\sum_{t=1}^{n}\varepsilon_t\varpi_t(K_0+1)I_t,
\end{aligned} \tag{5.29}$$

where $\varpi_t(K_0+1) = \mathbf{X}_{t,K_0+1}^{\intercal}\left(\tilde{\boldsymbol{\alpha}}_{K_0+1}(U_t) - \boldsymbol{\alpha}_{K_0+1}^*(U_t)\right)$. By some standard arguments and using (5.28), we may show that

$$\frac{1}{n_h}\sum_{t=1}^{n}\varepsilon_t^2 I_t = \sigma^2 + o_P(1), \tag{5.30}$$

$$\frac{1}{n_h}\sum_{t=1}^{n}\varpi_t^2(K_0+1)I_t = O_P\left(p^4\epsilon_n^2\right), \tag{5.31}$$

$$\begin{aligned}\frac{2}{n_h}\sum_{t=1}^{n}\varepsilon_t\varpi_t(K_0+1)I_t &= O_P\left(p^2(nh)^{-1} + p^2n^{-1}h^{-1/2} + p^2n^{-1/2}h^2\right)\\ &= o_P\left(p^4\epsilon_n^2\right),\end{aligned} \tag{5.32}$$

where the condition $nh^6 = o(1)$ in Assumption 7 is used in proving (5.32). By (5.29)–(5.32), we have

$$\begin{aligned}\mathsf{IC}(K_0+1) &= \log\left[\tilde{\sigma}_n^2(K_0+1)\right] + (K_0+1)\cdot\left[\frac{\log(nh)}{nh}\right]^{\rho}\\ &= \log(\sigma^2) + (K_0+1)\cdot\left[\frac{\log(nh)}{nh}\right]^{\rho} + O_P(p^4\epsilon_n^2)\\ &= \log(\sigma^2) + (K_0+1)\cdot\left[\frac{\log(nh)}{nh}\right]^{\rho}(1+o_P(1))\end{aligned} \tag{5.33}$$

as $p = o\left([\log(nh)/(nh)]^{(\rho-1)/4}\right)$ in Assumption 7. Similarly, for any $K_0 \le K \le \bar{K}$ and n sufficiently large, we can also prove (5.25). Details are omitted here to save the space. □

Proposition 2. *Suppose that the conditions of Theorem 2 are satisfied. For*

$1 \leq K \leq K_0 - 1$, *conditional on* $\mathbf{C}_n(K_0)$, *when* n *is sufficiently large,*

$$\mathsf{IC}(K) > \log(\sigma^2 + c_\alpha) + o_P(1) \tag{5.34}$$

and $\mathsf{IC}(K_0) = \log(\sigma^2) + o_P(1)$, *where* c_α *is defined in Assumption 5.*

Proof. The result of $\mathsf{IC}(K_0) = \log(\sigma^2) + o_P(1)$ can proved by using Proposition 1 with $K = K_0$. Hence, we only prove (5.34) for the case of $1 \leq K \leq K_0 - 1$. As discussed in Section 3, in this case, conditional on $\mathbf{C}_n(K_0)$, two or more clusters of $\mathcal{C}_k^0$, $k = 1, \cdots, K_0$, are falsely merged, which results in K clusters denoted by $\mathcal{C}_{1|K}, \cdots, \mathcal{C}_{K|K}$, respectively. Define $\mathbf{X}_{t,K}$ and the quasi functional coefficients $\boldsymbol{\alpha}_K(u)$ for the misspecified model as in (3.5) and (3.3), respectively. For notational simplicity, we next only consider the case of $K = K_0 - 1$. Other cases can be similarly handled but with slightly more complicated notation. Without loss of generality, we assume that the clusters $\mathcal{C}_{K_0-1}^0$ and $\mathcal{C}_{K_0}^0$ are first (falsely) merged, which indicates that

$$X_{t,k|K_0-1} = X_{t,k|K_0} \;\; 1 \leq k \leq K_0 - 2, \;\; X_{t,K_0-1|K_0-1} = X_{t,K_0-1|K_0} + X_{t,K_0|K_0}.$$

Let

$$\boldsymbol{\alpha}_{K_0-1}^{\diamond}(\cdot) = \left[\alpha_{1|K_0-1}(\cdot), \cdots, \alpha_{K_0-1|K_0-1}(\cdot), \alpha_{K_0-1|K_0-1}(\cdot)\right]^{\intercal},$$

where $\alpha_{k|K_0-1}(\cdot)$ is defined in (3.3). Note that conditional on $\mathbf{C}_n(K_0)$, $\tilde{\mathbf{X}}_{t,K_0-1} = \mathbf{X}_{t,K_0-1}$ and

$$\begin{aligned} & Y_t - \tilde{\mathbf{X}}_{t,K_0-1}^{\intercal}\tilde{\boldsymbol{\alpha}}_{K_0-1}(U_t) \\ = \;& \varepsilon_t + \mathbf{X}_{t,K_0}^{\intercal}\left[\boldsymbol{\alpha}_0(U_t) - \boldsymbol{\alpha}_{K_0-1}^{\diamond}(U_t)\right] - \mathbf{X}_{t,K_0-1}^{\intercal}\left[\tilde{\boldsymbol{\alpha}}_{K_0-1}(U_t) - \boldsymbol{\alpha}_{K_0-1}(U_t)\right] \\ =: \;& \varepsilon_t + \varpi_{t1}(K_0-1) + \varpi_{t2}(K_0-1). \end{aligned} \tag{5.35}$$

To further simplify notation, we let $\varpi_{t1} = \varpi_{t1}(K_0-1)$, $\varpi_{t2}(K_0-1) = \varpi_{t2}$ and $I_t = \mathsf{I}(U_t \in \mathcal{U}_h)$. From (5.35), we have

$$\begin{aligned} & \sum_{t=1}^{n}\left[Y_t - \tilde{\mathbf{X}}_{t,K_0-1}^{\intercal}\tilde{\boldsymbol{\alpha}}_{K_0-1}(U_t)\right]^2 \mathsf{I}(U_t \in \mathcal{U}_h) \\ = \;& \sum_{t=1}^{n}\varepsilon_t^2 I_t + \sum_{t=1}^{n}\varpi_{t1}^2 I_t + \sum_{t=1}^{n}\varpi_{t2}^2 I_t + \\ & 2\Big(\sum_{t=1}^{n}\varepsilon_t\varpi_{t1}I_t + \sum_{t=1}^{n}\varepsilon_t\varpi_{t2}I_t + \sum_{t=1}^{n}\varpi_{t1}\varpi_{t2}I_t\Big). \end{aligned} \tag{5.36}$$

Using Assumption 5, we may show that

$$\frac{1}{n_h}\sum_{t=1}^{n}\varpi_{t1}^2 I_t > c_\alpha(1 + o_P(1)). \tag{5.37}$$

By Assumption 6 and following the argument in the proof of Proposition 1,

we have

$$\sum_{t=1}^{n} \varpi_{t2}^2 I_t = O_P\left(np^4\epsilon_n^2\right) = o_P(n), \quad \sum_{t=1}^{n} \varepsilon_t \varpi_{t2} I_t = o_P\left(np^4\epsilon_n^2\right) = o_P(n). \tag{5.38}$$

Furthermore, we can also prove that

$$\sum_{t=1}^{n} \varepsilon_t \varpi_{t1} I_t = O_P\left(pn^{1/2}\right) = o_P(n), \tag{5.39}$$

$$\sum_{t=1}^{n} \varpi_{t1} \varpi_{t2} I_t = O_P\left(np^3\epsilon_n\right) = o_P(n) \tag{5.40}$$

as $p = o\left(\epsilon_n^{-1/3}\right)$ in Assumption 7.

Using (5.30) and (5.36)–(5.53), we readily have

$$\mathsf{IC}(K_0 - 1) > \log(\sigma^2 + c_\alpha) + o_P(1). \tag{5.41}$$

Similarly, we can prove (5.41) for any $1 \leq K \leq K_0 - 2$, completing the proof of the proposition. □

Before proving Theorem 3, we first give a proposition on the mean integrated squared error for the penalised local linear estimation defined in

Section 2.3. Conditional on $\mathbf{C}_n(K_0)$ and $\tilde{K} = K_0$, we define

$$\widehat{\mathbf{A}}_n = \left(\widehat{\mathbf{a}}_1^\intercal, \cdots, \widehat{\mathbf{a}}_n^\intercal\right)^\intercal, \quad \widehat{\mathbf{a}}_t = [\widehat{\alpha}_1(U_t), \cdots, \widehat{\alpha}_{K_0}(U_t)]^\intercal;$$

$$\widehat{\mathbf{B}}_n = \left(\widehat{\mathbf{b}}_1^\intercal, \cdots, \widehat{\mathbf{b}}_n^\intercal\right)^\intercal, \quad \widehat{\mathbf{b}}_t = \left[\widehat{\alpha}_1'(U_t), \cdots, \widehat{\alpha}_{K_0}'(U_t)\right]^\intercal.$$

Let $\mathbf{A}_0$ and $\mathbf{B}_0$ be defined similarly to $\widehat{\mathbf{A}}_n$ and $\widehat{\mathbf{B}}_n$ but with $\widehat{\alpha}_k(\cdot)$ and $\widehat{\alpha}_k'(\cdot)$ replaced by $\alpha_k^0(\cdot)$ and $\alpha_k^{0\prime}(\cdot)$, respectively.

Proposition 3. *Suppose that the conditions of Theorem 3 are satisfied. Then, we have*

$$\frac{1}{n}\left\|\widehat{\mathbf{A}}_n - \mathbf{A}_0\right\|^2 = O_P\left(\frac{p^4}{nh}\right), \quad \frac{1}{n}\left\|\widehat{\mathbf{B}}_n - \mathbf{B}_0\right\|^2 = O_P\left(\frac{p^4}{nh^3}\right) \tag{5.42}$$

conditional on $\mathbf{C}_n(K_0)$ *and* $\tilde{K} = K_0$.

Proof. The proof is similar to the arguments used in (Wang and Xia 2009) and (Li et al., 2015). Let

$$\mathbf{U}_1 = \left(\mathbf{u}_{11}^\intercal, \cdots, \mathbf{u}_{1n}^\intercal\right)^\intercal, \quad \mathbf{U}_2 = \left(\mathbf{u}_{21}^\intercal, \cdots, \mathbf{u}_{2n}^\intercal\right)^\intercal,$$

where both $\mathbf{u}_{1t} = (u_{1t,1}, \cdots, u_{1t,K_0})^\intercal$ and $\mathbf{u}_{2t} = (u_{2t,1}, \cdots, u_{2t,K_0})^\intercal$ are K_0-

dimensional column vectors, $t = 1, \cdots, n$. Define

$$\mathcal{C}_n(C) = \left\{(\mathbf{U}_1, \mathbf{U}_2) : \ \|\mathbf{U}_1\|^2 + \|\mathbf{U}_2\|^2 = nC\right\},$$

where C is a positive constant which may be sufficiently large. For $(\mathbf{U}_1, \mathbf{U}_2) \in \mathcal{C}_n(C)$, conditional on $\mathbf{C}_n(K_0)$ and $\tilde{K} = K_0$, we observe that

$$\mathcal{Q}_n\big(\mathbf{A}_0+\gamma_n\mathbf{U}_1, \mathbf{B}_0+\gamma_n\mathbf{U}_2/h\big) - \mathcal{Q}_n(\mathbf{A}_0, \mathbf{B}_0) = \mathcal{I}_n(1)+\mathcal{I}_n(2)+\mathcal{I}_n(3), \quad (5.43)$$

where $\gamma_n = \sqrt{p^4/(nh)}$,

$$\begin{aligned}
\mathcal{I}_n(1) &= \mathcal{L}_n\big(\mathbf{A}_0 + \gamma_n\mathbf{U}_1, \mathbf{B}_0 + \gamma_n\mathbf{U}_2/h\big) - \mathcal{L}_n(\mathbf{A}_0, \mathbf{B}_0), \\
\mathcal{I}_n(2) &= \sum_{k=1}^{K_0} p'_{\lambda_1}\big(\|\tilde{A}_k\|\big)\left(\|A_k^0 + \gamma_n U_{1k}\| - \|A_k^0\|\right), \\
\mathcal{I}_n(3) &= \sum_{k=1}^{\tilde{K}} p'_{\lambda_2}\big(\tilde{D}_k\big)\left(\|hB_k^0 + \gamma_n U_{2k}\| - \|hB_k^0\|\right),
\end{aligned}$$

A_k^0 and B_k^0 are defined in Section 3, $U_{1k} = (u_{11,k}, \cdots, u_{1n,k})^\intercal$ and $U_{2k} = (u_{21,k}, \cdots, u_{2n,k})^\intercal$.

We next study $\mathcal{I}_n(i)$, $i = 1, 2, 3$, in turn. Conditional on $\mathbf{C}_n(K_0)$ and $\tilde{K} = K_0$, we note that $\tilde{\mathbf{X}}_{t,K_0} = \mathbf{X}_{t,K_0}$,

$$\mathcal{L}_n(\mathbf{A}_0, \mathbf{B}_0) = \frac{1}{nh}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\varepsilon_t + \mathbf{X}_{t,K_0}^\intercal \mathbf{d}_{ts}\right)^2 K_h\left(U_t, U_s\right),$$

and

$$\begin{aligned}\mathcal{L}_n(\mathbf{A}_0+\gamma_n\mathbf{U}_1,\mathbf{B}_0+\gamma_n\mathbf{U}_2/h) &= \frac{1}{nh}\sum_{s=1}^{n}\sum_{t=1}^{n}\Big[\varepsilon_t+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{d}_{ts}-\gamma_n\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s}\\ &\quad -\gamma_n\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}(U_t-U_s)/h\Big]^2 K_h\left(U_t,U_s\right),\end{aligned}$$

where $\mathbf{d}_{ts}=\boldsymbol{\alpha}_0(U_t)-\boldsymbol{\alpha}_0(U_s)-\boldsymbol{\alpha}_0'(U_s)(U_t-U_s)$. For $\mathcal{I}_n(1)$, we then have

$$\begin{aligned}\mathcal{I}_n(1) &= -\frac{2\gamma_n}{nh}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\varepsilon_t+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{d}_{ts}\right)\left[\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s}+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}(U_t-U_s)/h\right]K_h\left(U_t,U_s\right)\\ &\quad +\frac{\gamma_n^2}{nh}\sum_{s=1}^{n}\sum_{t=1}^{n}\left[\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s}+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}(U_t-U_s)/h\right]^2 K_h\left(U_t,U_s\right)\\ &=: \mathcal{I}_n(4)+\mathcal{I}_n(5). \qquad (5.44)\end{aligned}$$

Letting

$$\mathbb{U}_{ts}=\begin{bmatrix}1 & (U_t-U_s)/h\\ (U_t-U_s)/h & (U_t-U_s)^2/h^2\end{bmatrix}$$

and $\otimes$ be the Kronecker product, for $\mathcal{I}_n(5)$, we may show that

$$\begin{aligned}\mathcal{I}_n(5) &= \frac{\gamma_n^2}{nh}\sum_{s=1}^{n}(\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})\left[\sum_{t=1}^{n}\left(\mathbf{X}_{t,K_0}\mathbf{X}_{t,K_0}^{\intercal}\right)\otimes\mathbb{U}_{ts}K_h\left(U_t,U_s\right)\right](\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})^{\intercal}\\ &= \gamma_n^2\sum_{s=1}^{n}(\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})\left[\frac{1}{nh}\sum_{t=1}^{n}\left(\mathbf{X}_{t,K_0}\mathbf{X}_{t,K_0}^{\intercal}\right)\otimes\mathbb{U}_{ts}K_h\left(U_t,U_s\right)\right](\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})^{\intercal}\\ &= \gamma_n^2\sum_{s=1}^{n}(\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})\left[f_U(U_s)\boldsymbol{\Sigma}_{X|K_0}(U_s)\otimes\boldsymbol{\Sigma}_K+O_P\left(p^2h^2+p^2\epsilon_n\right)\right](\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})^{\intercal}\\ &\geq \gamma_n^2(\zeta_1+o_P(1))\left(\|\mathbf{U}_1\|^2+\|\mathbf{U}_2\|^2\right), \qquad (5.45)\end{aligned}$$

where ϵ_n is defined in the proof of Theorem 1, ζ_1 is a positive constant bounded away from zero, and $\mathbf{\Sigma}_K = \mathsf{diag}(1, \mu_2)$ with $\mu_j = \int u^j K(u)du$ for $j \geq 1$. Observe that

$$\begin{aligned}
& \sum_{s=1}^{n}\sum_{t=1}^{n} \left(\varepsilon_t + \mathbf{X}_{t,K_0}^{\intercal}\mathbf{d}_{ts}\right)\left[\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} + \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}(U_t - U_s)/h\right] K_h\left(U_t, U_s\right) \\
= & \sum_{s=1}^{n}\sum_{t=1}^{n} \varepsilon_t \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} K_h\left(U_t, U_s\right) + \sum_{s=1}^{n}\sum_{t=1}^{n} \varepsilon_t \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}\left((U_t - U_s)/h\right) K_h\left(U_t, U_s\right) + \\
& \sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{X}_{t,K_0}^{\intercal}\mathbf{d}_{ts}\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} K_h\left(U_t, U_s\right) + \\
& \sum_{s=1}^{n}\sum_{t=1}^{n} \mathbf{X}_{t,K_0}^{\intercal}\mathbf{d}_{ts}\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}\left((U_t - U_s)/h\right) K_h\left(U_t, U_s\right) \\
=: & \ \mathcal{I}_n(4,1) + \mathcal{I}_n(4,2) + \mathcal{I}_n(4,3) + \mathcal{I}_n(4,4).
\end{aligned} \tag{5.46}$$

Noting that the observations are independent as assumed in Assumption 3(ii), we have $\mathsf{E}\left[\mathcal{I}_n(4,1)\right] = 0$, and

$$\begin{aligned}
\mathsf{E}\left[\mathcal{I}_n^2(4,1)\right] &= \mathsf{E}\left[\left(\sum_{s=1}^{n}\sum_{t=1}^{n} \varepsilon_t \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} K_h\left(U_t, U_s\right)\right)^2\right] \\
&\leq n\sum_{s=1}^{n} \mathsf{E}\left[\left(\sum_{t=1}^{n} \varepsilon_t \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} K_h\left(U_t, U_s\right)\right)^2\right] \\
&= O\left(p^2 n^2 h\right) \cdot \|\mathbf{U}_1\|^2.
\end{aligned} \tag{5.47}$$

Similarly, we can also show that $\mathsf{E}\left[\mathcal{I}_n(4,2)\right] = 0$

$$\mathsf{E}\left[\mathcal{I}_n^2(4,2)\right] = O\left(p^2 n^2 h\right) \cdot \|\mathbf{U}_2\|^2. \tag{5.48}$$

By Taylor's expansion on the functional coefficients, we have

$$\mathsf{E}\left[|\mathcal{I}_n(4,3)|\right] = O\left(p^2 n^{3/2} h^3\right) \cdot \|\mathbf{U}_1\| \tag{5.49}$$

and

$$\mathsf{E}\left[|\mathcal{I}_n(4,4)|\right] = O\left(p^2 n^{3/2} h^3\right) \cdot \|\mathbf{U}_2\|. \tag{5.50}$$

Following (5.47)–(5.50) and noting that $p^2 n h^5 = O(1)$ in Assumption 9, we may show that

$$\mathcal{I}_n(4) = O_P\left(\gamma_n^2 n^{1/2}\right) \cdot \left(\|\mathbf{U}_1\| + \|\mathbf{U}_2\|\right). \tag{5.51}$$

By choosing the constant C sufficiently large, $\mathcal{I}_n(4)$ would be asymptotically dominated by $\mathcal{I}_n(5)$. As a result, we have

$$\mathcal{I}_n(1) \geq \gamma_n^2(\zeta_1/2 + o_P(1)) \left(\|\mathbf{U}_1\|^2 + \|\mathbf{U}_2\|^2\right). \tag{5.52}$$

We next consider $\mathcal{I}_n(2)$. It is easy to see that

$$\begin{aligned} \mathcal{I}_n(2) &= \sum_{k=1}^{K_0} p'_{\lambda_1}\left(\|\tilde{A}_k\|\right) \left(\|A_k^0 + \gamma_n U_{1k}\| - \|A_k^0\|\right) \\ &\geq \sum_{k=1}^{K_0-1} p'_{\lambda_1}\left(\|\tilde{A}_k\|\right) \left(\|A_k^0 + \gamma_n U_{1k}\| - \|A_k^0\|\right) \end{aligned} \tag{5.53}$$

as $\|A^0_{K_0}\| = 0$. Furthermore, following the argument in the proof of Theorem 2, we may show that

$$\|\tilde{A}_k\| = \|A^0_k\| + O_P\left(n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n\right) = \|A^0_k\| + o_P(n^{1/2}),$$

which together with Assumption 8, indicates that

$$\|\tilde{A}_k\| \geq c_A\sqrt{n}/2$$

with probability approaching one. By the definition of the SCAD penalty derivative and noting that $\lambda_1 = o(n^{1/2})$ in (3.9), we have $\mathcal{I}_n(2) \geq 0$ with probability approaching one. Analogously, we can also show that $\mathcal{I}_n(3) \geq 0$ with probability approaching one. Hence, for any small $\epsilon > 0$ there exists sufficiently large $C > 0$ such that

$$\mathsf{P}\Big\{\inf_{(\mathbf{U}_1,\mathbf{U}_2)\in\mathcal{C}_n(C)} \mathcal{Q}_n\big(\mathbf{A}_0 + \gamma_n\mathbf{U}_1, \mathbf{B}_0 + \gamma_n\mathbf{U}_2/h\big) > \mathcal{Q}_n(\mathbf{A}_0, \mathbf{B}_0)\Big\} \geq 1 - \epsilon \tag{5.54}$$

for large n, which leads to (5.42), completing the proof of this proposition.□

Proof of Theorem 3. Letting $\tilde{\mathbf{C}}_n = \mathbf{C}_n(K_0) \cap \{\tilde{K} = K_0\}$, observe that

$$\mathsf{P}\left(\|\widehat{A}_{K_0}\| = 0\right) = \mathsf{P}\left(\|\widehat{A}_{K_0}\| = 0|\tilde{\mathbf{C}}_n\right)\mathsf{P}\left(\tilde{\mathbf{C}}_n\right) + \mathsf{P}\left(\|\widehat{A}_{K_0}\| = 0|\tilde{\mathbf{C}}_n^c\right)\mathsf{P}\left(\tilde{\mathbf{C}}_n^c\right), \tag{5.55}$$

which together with Theorems 1 and 2, implies that

$$\mathsf{P}\left(\|\widehat{A}_{K_0}\| = 0|\tilde{\mathbf{C}}_n\right) \to 1 \tag{5.56}$$

is sufficient for our proof. Recall that $X_{t,k|K_0} = \sum_{j\in\mathcal{C}_k^0} X_{tj}$ and define $\mathcal{L}'_{nk,1}(\mathbf{A}, \mathbf{B})$ be an n-dimensional vector with the s-th component being

$$\mathcal{L}'_{nk,1s} = \frac{2}{n}\sum_{t=1}^{n} X_{t,k|K_0}\left[Y_t - \mathbf{X}_{t,K_0}^{\intercal}\mathbf{a}_s - \mathbf{X}_{t,K_0}^{\intercal}\mathbf{b}_s(U_t - U_s)\right]K_h(U_t, U_s).$$

When $\|A_k\| \neq 0$, let $\mathcal{P}'_{n1}(A_k)$ be an n-dimensional vector with the s-th component being

$$\mathcal{P}'_{n1,s}(A_k) = p'_{\lambda_1}\big(\|\tilde{A}_k\|\big)\frac{a_{sk}}{\|A_k\|}.$$

Following the arguments in the proof of Theorem 2 above, we may show that

$$\|\tilde{A}_{K_0}\| = \|A^0_{K_0}\| + O_P\left(n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n\right) = O_P\left(n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n\right) = o_P(\lambda_1). \tag{5.57}$$

From the definition of $p'_{\lambda_1}(\cdot)$ and (5.57), when $\|\widehat{A}_{K_0}\| \neq 0$, we have

$$\|\mathcal{P}'_{n1}(A_{K_0})\| = \lambda_1 \tag{5.58}$$

with probability approaching one. If $\|\widehat{A}_{K_0}\| \neq 0$, we must have

$$\mathcal{L}'_{nk,1}(\widehat{\mathbf{A}}_n, \widehat{\mathbf{B}}_n) = \mathcal{P}'_{n1}(\widehat{A}_k) \tag{5.59}$$

for $k = K_0$. However, using Proposition 3, we can prove that

$$\left\| \mathcal{L}'_{nk,1}(\widehat{\mathbf{A}}_n, \widehat{\mathbf{B}}_n) \right\| = O_P\left(n^{1/2} p\epsilon_n + p^4/h^{1/2}\right) = o_P(\lambda_1),$$

which together with (5.58), indicates that (5.59) cannot hold. Therefore, conditional on $\tilde{\mathbf{C}}_n$, $\|\widehat{A}_{K_0}\|$ must be zero with probability approaching one.

Similarly, we can also prove that

$$\mathsf{P}\left(\|\widehat{B}_k\| = 0,\ k = K_*, \cdots, K_0\right) \to 1. \tag{5.60}$$

The proof of Theorem 3 has been completed. □

Chapter 6

Conclusions and Future Work

In this thesis, the kernel-based hierarchical clustering method and a generalised version of information criterion have been developed to uncover the latent homogeneity structure in the classic functional-coefficient models. Furthermore, the penalised local linear estimation approach is used to separate out the zero-constant cluster, the non-zero constant-coefficient clusters and the functional-coefficient clusters. The asymptotic theory in Chapter 5 shows that the estimation for the true number of clusters and the true set of clusters is consistent in the large-sample case. In the simulation study, we find that the proposed estimation methodology outperforms the direct nonparametric kernel estimation which ignores the latent structure in the model. In empirical application to the Boston house price data and plasma beta-carotene level data, we show that the nonparametric functional-

coefficient model can be substantially simplified with reduced numbers of unknown parametric and nonparametric components. As a result, the out-of-sample mean squared prediction errors using the developed approach are significantly smaller than those using the naive kernel method which ignores the latent homogeneity structure among the functional coefficients.

In the future, we may further do some extensions in terms of generalised varying coefficient model and combine it with proposed homogeneity structure.

References

Bellman, R. (1957). *Dynamic programming*. Princeton: Princeton University Press.

Bondell, H. D. and B. J. Reich (2008). Simultaneous regression shrinkage, variable selection and supervised clustering of predictors with oscar. *Biometrics 64*, 115–123.

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

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

Chen, J., D. Li, and Y. Xia (2018). Estimation of a rank-reduced functional-

coefficient panel data model in presence of serial correlation. *Working paper*.

Chen, R. and R. Tsay (1993). Functional-coefcient autoregressive models. *Journal of the American Statistical Association 88*, 298–308.

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

Cleveland, W., E. Grosse, and W. Shyu (1991). Local regression models. *In Statistical Models in S*, 309–376.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011). *Cluster analysis* (5th ed.). Wiley: Wiley Series in Probability and Statistics.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *The Annals of Statistics 21*, 196–216.

Fan, J. and I. Gijbels (1996). *Local polynomial modelling and its applications*. London: Chapman and Hall.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized

likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

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

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

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

Fan, Y. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association 87*, 998–1004.

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

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

Friedman, J., T. Hastie, H. Hofling, and R. Tibshirani (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics 1*, 302—332.

Green, P. and B. Silverman (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall/CRC.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient model. *Journal of the Royal Statistical Socity. Series B 55*, 757–796.

Jiang, Q., H. Wang, Y. Xia, and G. Jiang (2013). On a principal varying coefficient model. *Journal of the American Statistical Association 108*, 228–236.

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

Ke, Y., J. Li, and W. Zhang (2016). Structure identification in panel data analysis. *The Annals of Statistics 44*, 1193–1233.

Ke, Z., J. Fan, and Y. Wu (2015). Homogeneity pursuit. *Journal of the American Statistical Association 110*, 175–194.

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

Liu, J., R. Li, and R. Wu (2014). Feature selection for varying coefficient

models with ultrahigh dimensional covariates. *Journal of the American Statistical Association 109*, 266–274.

Mack, Y. P. and B. W. Silverman (1982). Weak and strong uniform consistency for kernel regression estimates. *Zeitschrift fur Wahrscheinlichskeitheorie und verwandte Gebiete 61*, 405–415.

McCullagh, P. and J. Nelder FRS (1989). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman and Hall.

Nierenberg, D., T. Stukel, J. Baron, B. Dain, and E. Greenberg (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology 130*, 511–521.

Park, B., E. Mammen, Y. K. Lee, and E. Lee (2015). Varying coefficient regression models: a review and new developments. *International Statistical Review 83*, 36–64.

Rencher, A. C. and W. F. Christensen (2012). *Methods of multivariate analysis* (3rd ed.). Wiley: Wiley Series in Probability and Statistics.

Rupper, D., S. Sheather, and M. a. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*, 1257–1270.

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

Shen, X. and H. C. Huang (2010). Group pursuit through a regularization solution surface. *Journal of the American Statistical Association 105*, 727–739.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society, SeriesB 36*, 111–147.

Su, L., Z. Shi, and P. C. B. Phillips (2016). Identifying latent structures in panel data. *Econometrica 84*, 2215–2264.

Su, L., X. Wang, and S. Jin (2017). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business and Economic Statistics*.

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

Tibshirani, R., M. Saunders, S. Rosset, and K. Knight (2005). Sparsity and smoothness visa the fused lasso. *Journal of Royal Satistical Society, Serie B 67*, 91–108.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005).

Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B 67*, 91–108.

Vogt, M. and O. Linton (2017). Classification of nonparametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society, Series B 79*, 5–27.

Wand, M. P. and M. C. Jones (1994). *Kernel smoothing*. Chapman and Hall.

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

Wang, L. and R. Li (2009). Weighted wilcoxon-type smoothly clipped absolute deviation method. *Biometrics 65*, 564–571.

Xia, Y., W. Zhang, and H. Tong (2004). Efficient estimation for semivarying-coefficient models. *Biometrika 91*, 661–681.

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

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