UNIVERSITY of York

Discussion Papers in Economics

No. 19/03

Nonparametric Homogeneity Pursuit in Functional-Coefficient Models

Jia Chen, Degui Li, Lingling Wei, Wenyang Zhang

Department of Economics and Related Studies
University of York
Heslington
York, YO10 5DD

Nonparametric Homogeneity Pursuit in Functional-Coefficient Models

Jia Chen[1], Degui Li[2], Lingling Wei[2], Wenyang Zhang[2]

[1]Department of Economics and Related Studies, University of York, United Kingdom

[2]Department of Mathematics, University of York, United Kingdom

9 February 2019

Abstract

This paper explores the homogeneity of coefficient functions in nonlinear models with functional coefficients and identifies the underlying semiparametric modelling structure. With initial kernel estimates of coefficient functions, we combine the classic hierarchical clustering method with a generalised version of the information criterion to estimate the number of clusters, each of which has a common functional coefficient, and determine the membership of each cluster. To identify a possible semi-varying coefficient modelling framework, we further introduce a penalised local least squares method to determine zero coefficients, non-zero constant coefficients and functional coefficients which vary with an index variable. Through the nonparametric kernel-based cluster analysis and the penalised approach, we can substantially reduce the number of unknown parametric and nonparametric components in the models, thereby achieving the aim of dimension reduction. Under some regularity conditions, we establish the asymptotic properties for the proposed methods including the consistency of the homogeneity pursuit. Numerical studies, including Monte-Carlo experiments and an empirical application, are given to demonstrate the finite-sample performance of our methods.

Keywords: Functional-coefficient models, Hierarchical clustering, Homogeneity, Information criterion, Nonparametric estimation, Penalised method.

1 Introduction

We consider the functional-coefficient model defined by

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

where Y_t is a response variable, $\mathbf{X}_t = (X_{t1}, \cdots, X_{tp})^\intercal$ is a p-dimensional vector of random covariates, $\boldsymbol{\beta}_0(\cdot) = \left[\beta_1^0(\cdot), \cdots, \beta_p^0(\cdot)\right]^\intercal$ is a p-dimensional vector of functional coefficients, U_t is a univariate index variable, and ε_t is an independent and identically distributed (*i.i.d.*) error term. The functional-coefficient model is a natural extension of the classic linear regression model by allowing the regression coefficients to vary with certain index variable, and thus captures flexible dynamic relationship between the response and covariates. In recent years, there have been extensive studies on estimation and model selection for model (1.1) and its various generalised versions, see, for example, 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 (1.1) would be unstable. To address this issue, there have been some extensive studies in the literature on 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 aim to reduce the dimension of unknown functional coefficients and improve estimation efficiency. In this paper we consider a different approach, i.e., we assume that there is a homogeneity structure on model (1.1) so that individual functional coefficients can be grouped into a number of clusters and coefficients within each cluster have the same functional pattern. Throughout the paper, we assume that the dimension p may depend on the sample size n and can be divergent with n, but the number of unknown clusters is 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 coefficients. Specifically, we assume the following homogeneity structure on model (1.1): there exists a partition of $\{1, 2, \cdots, p\}$ denoted by $\mathcal{C}_0 = \left\{\mathcal{C}_1^0, \cdots, \mathcal{C}_{K_0}^0\right\}$ such that

$$\beta_j^0(\cdot) = \alpha_k^0(\cdot) \ \text{ for } \ j \in \mathcal{C}_k^0 \ \text{ and } \ \mathcal{C}_{k_1}^0 \cap \mathcal{C}_{k_2}^0 = \emptyset \ \text{ for } \ 1 \leqslant k_1 \neq k_2 \leqslant K_0, \tag{1.2}$$

where the Lebesgue measure of $\left\{u \in \mathcal{U} : \alpha_{k_1}^0(u) - \alpha_{k_2}^0(u) \neq 0\right\}$ is positive and bounded away from zero for any $1 \leqslant k_1 \neq k_2 \leqslant K_0$, 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, in which case model (1.1) is semiparametric with a combination of constant and functional coefficients. Our aim is to (i) explore the homogeneity structure (1.2) by estimating the *unknown* number of clusters K_0 and identifying members of the clusters $\mathcal{C}_1^0, \cdots, \mathcal{C}_{K_0}^0$; and (ii) identify the clusters of constant coefficients and those of coefficients varying with U_t and estimate their *unknown* values.

The topic investigated in our paper has two close relatives in existing literature. On one hand, the functional-coefficient regression with the homogeneity structure is a natural extension of linear regression with homogeneity structure, which has received increasing attention in recent years. For example, Tibshirani *et al* (2005) introduce the so-called fused LASSO method to study 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, this paper 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. The methodology of nonparametric homogeneity pursuit developed in this paper will be substantially different from those in the aforementioned literature.

In this paper, we first estimate each functional coefficient in model (1.1) by using the kernel smoothing method and ignoring the homogeneity structure (1.2), and calculate the L_1-distance 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 (1.2), 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 (1.2), we further estimate a semi-varying coefficient modelling framework by determining the zero coefficients, 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 as derivatives of the well-known SCAD penalty introduced by Fan and Li (2001). With the nonparametric cluster analysis and the penalised approach, we can reduce the number of unknown components in model (1.1) from p to $K_0 - 1$ (if the zero constant coefficients exist in the model). Furthermore, the choice of the tuning parameters in the proposed estimation approach and the computational algorithm is also discussed. 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 and find that the original nonparametric functional-coefficient models can be simplified and the number of unknown components involved can be reduced. In particular, the out-of-sample mean absolute prediction errors of our approach are usually much smaller than those using the naive kernel method which ignores the latent homogeneity structure.

The rest of the paper is organised as follows. Section 2 introduces the clustering method, information criterion and penalised method to determine the unknown clusters and estimate the unknown components. Section 3 establishes the asymptotic theory for the proposed clustering and estimation methods. Section 4 discusses the choice of the tuning parameters and introduces an algorithm for computing the penalised estimates. Section 5 reports a Monte-Carlo simulation study. Section 6 gives the empirical application to the Boston house price data. Section 7 concludes the paper. The proofs of the main asymptotic theorems are given in a supplemental document.

2 Methodology

In this section, we first introduce a clustering method for kernel estimated functional coefficients in Section 2.1, followed by a generalised information criterion for determining the number of clusters in Section 2.2, and finally propose a penalised local linear estimation approach to identify the semi-varying coefficient modelling structure in Section 2.3.

2.1 Kernel-based clustering method

Assuming that the coefficient functions have continuous second-order derivatives, we can use the kernel smoothing method (c.f., Wand and Jones, 1994) to obtain preliminary estimates of $\beta_j^0(\cdot)$, $j = 1, \cdots, p$, which are denoted by $\tilde{\beta}_j(\cdot)$, $j = 1, \cdots, p$. 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

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}(u_0) &= \left[\tilde{\beta}_1(u_0), \cdots, \tilde{\beta}_p(u_0)\right]^\intercal \\
&= \left[\sum_{t=1}^{n} \mathbf{X}_t \mathbf{X}_t^\intercal K_h(U_t, u_0)\right]^{-1} \left[\sum_{t=1}^{n} \mathbf{X}_t Y_t K_h(U_t, u_0)\right] \\
&= \left[\mathbb{X}_n^\intercal \mathbb{W}_n(u_0) \mathbb{X}_n\right]^{-1} \left[\mathbb{X}_n^\intercal \mathbb{W}_n(u_0) \mathbb{Y}_n\right], \qquad (2.1)
\end{aligned}$$

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

$$\tilde{\Delta}_{ij} = \frac{1}{n}\sum_{t=1}^{n} \left|\tilde{\beta}_i(U_t) - \tilde{\beta}_j(U_t)\right| I(U_t \in \mathcal{U}_h), \tag{2.2}$$

where $I(\cdot)$ 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

$$\Delta_{ij}^0 = \int_{\mathcal{U}_h} \left|\beta_i^0(u) - \beta_j^0(u)\right| f_U(u)du, \tag{2.3}$$

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

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

From (1.2) and (2.3), 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, denoted by $\mathbf{\Delta}_0$, whose (i, j)-entry is Δ_{ij}^0. The corresponding estimated distance matrix, denoted by $\tilde{\mathbf{\Delta}}_n$, has entries $\tilde{\Delta}_{ij}$ defined in (2.2). 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 singleton clusters, corresponding to the p functional coefficients. In each stage, the two clusters with the smallest distance are merged into a new cluster. This continues until we end with only one full cluster. Such a clustering approach has been widely studied in the literature of cluster analysis (c.f., Everitt *et 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 paper fills in this gap. Specifically, the algorithm is described as follows, where the number of clusters K_0 is assumed to be known. Section 2.2 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 farthest distance between a point in $\mathcal{A}$ and a point in $\mathcal{B}$, which is called the complete linkage.*

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 $1 \leqslant K \leqslant p$, 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.

2.2 Estimation of the cluster number

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

$$\begin{aligned}\tilde{\boldsymbol{\alpha}}_K(u_0) &= \left[\tilde{\alpha}_{1|K}(u_0), \cdots, \tilde{\alpha}_{K|K}(u_0)\right]^\intercal \\ &= \left[\sum_{t=1}^n \tilde{\mathbf{X}}_{t,K}\tilde{\mathbf{X}}_{t,K}^\intercal K_h(U_t, u_0)\right]^{-1}\left[\sum_{t=1}^n \tilde{\mathbf{X}}_{t,K} Y_t K_h(U_t, u_0)\right],\end{aligned} \tag{2.4}$$

where

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

$\tilde{\mathcal{C}}_{k|K}$ is defined as in Section 2.1. 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 the Appendix); but when K is smaller than K_0, the clustering approach in Section 2.1 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 (3.3) below.

We define the following objective function:

$$\mathsf{IC}(K) = \log\left[\tilde{\sigma}_n^2(K)\right] + K \cdot \left[\frac{\log(nh)}{nh}\right]^\rho \tag{2.5}$$

with $0 < \rho < 1$,

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

and determine the number of clusters through

$$\tilde{K} = \arg \min_{1 \leqslant K \leqslant \bar{K}} \mathsf{IC}(K), \tag{2.6}$$

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. 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 Schwarz (1978). The latter has been extended to the nonparametric framework in recent years (c.f., Wang and Xia, 2009).

2.3 Penalised local linear estimation

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 $\tilde{\boldsymbol{\alpha}}_K(u_0)$ with $K = \tilde{K}$. Throughout the paper, we call $\tilde{\boldsymbol{\alpha}}(\cdot)$ the *post-clustering kernel estimator*. 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, is zero (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 estimate the 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

$$\mathbf{A} = \left(\mathbf{a}_1^\intercal, \cdots, \mathbf{a}_n^\intercal\right)^\intercal, \ \mathbf{a}_t = (a_{t1}, \cdots, a_{t\tilde{K}})^\intercal;$$
$$\mathbf{B} = \left(\mathbf{b}_1^\intercal, \cdots, \mathbf{b}_n^\intercal\right)^\intercal, \ \mathbf{b}_t = (b_{t1}, \cdots, b_{t\tilde{K}})^\intercal;$$
$$A_k = (a_{1k}, \cdots, a_{nk})^\intercal, \ B_k = (b_{1k}, \cdots, b_{nk})^\intercal.$$

We define the penalised objective function as follows:

$$\mathcal{Q}_n(\mathbf{A},\mathbf{B}) = \mathcal{L}_n(\mathbf{A},\mathbf{B}) + \mathcal{P}_{n1}(\mathbf{A}) + \mathcal{P}_{n2}(\mathbf{B}), \tag{2.7}$$

where

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

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):

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

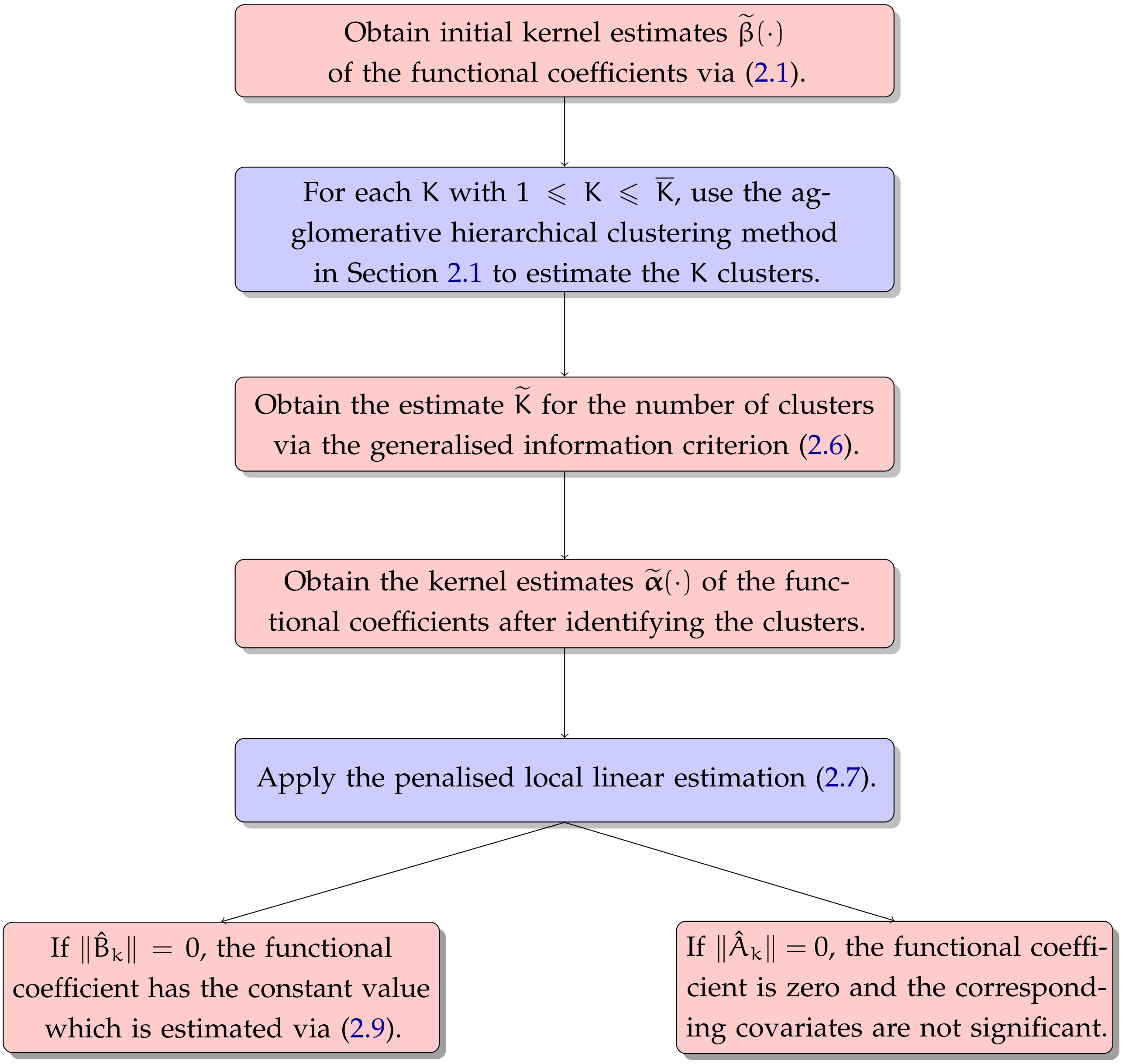
Let

$$\widehat{A}_k = [\widehat{\alpha}_k(U_1),\ \cdots,\ \widehat{\alpha}_k(U_n)]^{\intercal} \ \text{ and } \ \widehat{B}_k = [\widehat{\alpha}'_k(U_1),\ \cdots,\ \widehat{\alpha}'_k(U_n)]^{\intercal},\ \ k=1,\cdots,\tilde{K}, \tag{2.8}$$

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 with zero coefficient, and $\|\widehat{B}_k\| = 0$ when $\tilde{\mathcal{C}}_{k|\tilde{K}}$ is the estimated cluster with a non-zero constant coefficient. Hence, if $\|\widehat{A}_k\| = 0$, the corresponding covariates are not significant and should be removed from the functional-coefficient model (1.1); and if $\|\widehat{B}_k\| = 0$, the functional coefficient has a constant value and can be consistently estimated by

$$\widehat{\alpha}_k = \frac{1}{n}\sum_{t=1}^{n}\widehat{\alpha}_k(U_t). \tag{2.9}$$

Implementation of the proposed methods in Sections 2.1–2.3 is summarised in the following flowchart.


Obtain initial kernel estimates $\widetilde{\beta}(\cdot)$ of the functional coefficients via (2.1).
For each K with $1 \leqslant K \leqslant \overline{K}$, use the agglomerative hierarchical clustering method in Section 2.1 to estimate the K clusters.
Obtain the estimate $\widetilde{K}$ for the number of clusters via the generalised information criterion (2.6).
Obtain the kernel estimates $\widetilde{\boldsymbol{\alpha}}(\cdot)$ of the functional coefficients after identifying the clusters.
Apply the penalised local linear estimation (2.7).
If $\|\hat{B}_k\| = 0$, the functional coefficient has the constant value which is estimated via (2.9).
If $\|\hat{A}_k\| = 0$, the functional coefficient is zero and the corresponding covariates are not significant.

Flowchart for implementing the methods proposed in Sections 2.1–2.3.

3 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,*

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

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 $\mathsf{E}\left(|\varepsilon_t|^{2+\iota_1}\right) + \max_{1\leqslant i\leqslant p} \mathsf{E}\left(|X_{ti}|^{2(2+\iota_1)}\right) < \infty$.*

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

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

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 $0 \leqslant \delta_0 < 2/5$, 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*

$$\mathsf{P}\left(\left\{\tilde{\mathcal{C}}_k,\ k=1,\cdots,K_0\right\}\neq\left\{\mathcal{C}_k^0,\ k=1,\cdots,K_0\right\}\right)=o(1) \tag{3.1}$$

when the sample size n is sufficiently large, where $\tilde{\mathcal{C}}_k$ is defined in Section 2.1 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 2.1 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

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

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:

$$\mathbf{C}_n(K_0)=\left\{\left[\tilde{\mathcal{C}}_k,\ k=1,\cdots,K_0\right]=\left[\mathcal{C}_k^0,\ k=1,\cdots,K_0\right]\right\}. \tag{3.2}$$

From (3.1) 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, $1\leqslant K\leqslant K_0-1$. With such a clustering result, the functional coefficients in model (1.1) and (1.2) cannot be consistently estimated by the kernel smoothing method, as the model is misspecified. However, we may define the "*quasi*" functional coefficients by

$$\boldsymbol{\alpha}_K(u)=\left[\alpha_{1|K}(u),\cdots,\alpha_{K|K}(u)\right]^\intercal=\left[\mathbf{\Sigma}_{X|K}(u)\right]^{-1}\mathbf{\Sigma}_{XY|K}(u), \tag{3.3}$$

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

$$\mathbf{\Sigma}_{X|K}(u)=\mathsf{E}\left[\mathbf{X}_{t,K}\mathbf{X}_{t,K}^\intercal|U_t=u\right],\ \ \mathbf{\Sigma}_{XY|K}(u)=\mathsf{E}\left[\mathbf{X}_{t,K}Y_t|U_t=u\right], \tag{3.4}$$

and

$$\mathbf{X}_{t,K} = \left(X_{t,1|K}, \cdots, X_{t,K|K}\right)^{\intercal} \quad \text{with } X_{t,k|K} = \sum_{j \in \mathcal{C}_{k|K}} X_{tj}. \tag{3.5}$$

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 (3.3), it is easy to show that

$$\mathsf{E}\left[\boldsymbol{\varepsilon}_{t1,K}|U_t\right] = \mathbf{0} \ \text{ a.s.}, \tag{3.6}$$

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 (2.4) of Section 2.2, where the order of elements in the latter may have to be re-arranged if necessary. The fact of (3.6) 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 the supplemental document).

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 2.2.

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

$$\inf_{\boldsymbol{\alpha}(\cdot) \in \mathcal{A}(K_0)} \int_0^1 \left[\boldsymbol{\alpha}_0(u) - \boldsymbol{\alpha}(u)\right]^{\intercal} \boldsymbol{\Sigma}_{X|K_0}(u) \left[\boldsymbol{\alpha}_0(u) - \boldsymbol{\alpha}(u)\right] f_U(u) du > c_\alpha. \tag{3.7}$$

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

(ii). *For any* $1 \leqslant K \leqslant K_0$*, 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 (2.5).*

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 (3.7) in Assumption 5 indicates that the K_0 functional elements in $\boldsymbol{\alpha}_0(\cdot)$ needs to be "sufficiently" distinct. We may show that (3.7) is satisfied if $\inf_{1 \leqslant K \leqslant K_0} \inf_{u \in [0,1]} \lambda_{\min}\left(\boldsymbol{\Sigma}_{X|K}(u)\right) > c_1 > 0$ 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 $k_1 \neq k_2$. 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 (2.5) is consistent.

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

$$\mathsf{P}\left(\tilde{K} = K_0\right) \to 1, \tag{3.8}$$

where $\tilde{K}$ is defined in (2.6).

Define

$$A_k^0 = \left[\alpha_k^0(U_1), \cdots, \alpha_k^0(U_n)\right]^\intercal, \ \ B_k^0 = \left[\alpha_k^{0\prime}(U_1), \cdots, \alpha_k^{0\prime}(U_n)\right]^\intercal,$$
$$\widehat{A}_k = [\widehat{\alpha}_k(U_1), \cdots, \widehat{\alpha}_k(U_n)]^\intercal, \ \ \widehat{B}_k = [\widehat{\alpha}_k^\prime(U_1), \cdots, \widehat{\alpha}_k^\prime(U_n)]^\intercal.$$

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 Sections 2.1 and 2.2 is applied to the penalised local linear estimation. Some additional conditions are needed to derive the sparsity result for the penalised estimation in Section 2.3.

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

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

$$\lambda_1 = o(n^{1/2}), \ \ n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n + p^4h^{-1/2} = o(\lambda_1). \tag{3.9}$$

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 (2.7) 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 $0.1 < \delta_* < 0.5$, (3.9) in Assumption 9 would be satisfied.

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

$$\mathsf{P}\left(\|\widehat{A}_{K_0}\| = 0,\ \|\widehat{B}_k\| = 0, k = K_*, \cdots, K_0 \right) \to 1. \tag{3.10}$$

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.

4 Practical issues in the estimation procedure

In this section, we first discuss how to choose the bandwidth in the kernel estimation and the tuning parameters in the penalised local least squares estimation; we then introduce an easy-to-implement computational algorithm for the penalised approach in Section 2.3.

4.1 Choice of tuning parameters

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

$$\mathsf{CV}(h) = \frac{1}{n} \sum_{t=1}^{n} \left[Y_t - \mathbf{X}_t^\intercal \tilde{\boldsymbol{\beta}}_{-t}(U_t|h) \right]^2, \tag{4.1}$$

where $\tilde{\boldsymbol{\beta}}_{-t}(\cdot|h)$ is the preliminary kernel estimator of $\boldsymbol{\beta}_0(\cdot)$ in model (1.1) using the bandwidth h and all observations except the t-th observation. Then we determine the optimal bandwidth $\hat{h}_{opt}$ by minimising $\mathsf{CV}(h)$ with respect to h. The cross-validation criterion for bandwidth selection in the post-clustering kernel estimation $\widetilde{\boldsymbol{\alpha}}(\cdot)$ can be defined in exactly the same way.

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 denote $\mathcal{M}_1(\boldsymbol{\lambda})$ and $\mathcal{M}_2(\boldsymbol{\lambda})$ 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:

$$\begin{aligned}\mathsf{GIC}(\boldsymbol{\lambda}) &= \sum_{t=1}^{n}\left[Y_t - \sum_{k\in\mathcal{M}_1(\boldsymbol{\lambda})} \tilde{X}_{t,k|\tilde{K}}\widehat{\alpha}_{k,\lambda}(U_t) - \sum_{k\in\mathcal{M}_2(\boldsymbol{\lambda})} \tilde{X}_{t,k|\tilde{K}}\widehat{\alpha}_{k,\lambda}\right]^2 \\ &\quad +2\ln[\ln(n)]\ln(m_0h^{-1})(|\mathcal{M}_2(\boldsymbol{\lambda})| + |\mathcal{M}_1(\boldsymbol{\lambda})|m_0h^{-1}), \qquad (4.2)\end{aligned}$$

where $\widehat{\alpha}_{k,\lambda}(\cdot)$ and $\widehat{\alpha}_{k,\lambda}$ are defined as the penalised estimation in Section 2.3 using the tuning parameter vector $\boldsymbol{\lambda}$, $|\mathcal{M}|$ denotes the cardinality of the set $\mathcal{M}$, and the bandwidth h can be determined by the leave-one-out cross-validation. The optimal value of $\boldsymbol{\lambda}$ can be found by minimising the objective function $\mathsf{GIC}(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$.

4.2 Computational algorithm for penalised estimation

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 $\tilde{\Omega}_{nk,s}(j) = \frac{2}{nh}\sum_{t=1}^{n}\tilde{X}_{t,k|\tilde{K}}\tilde{X}_{t,k|\tilde{K}}\left[(U_t - U_s)/h\right]^j K_h(U_t, U_s)$. We next introduce an iterative procedure to compute the penalised local least squares estimates of the functional coefficients proposed in Section 2.3 (see Li, Ke and Zhang, 2015). It can be viewed as a nonparametric extension of the coordinate descent algorithm, which is a commonly used optimisation algorithm that finds the minimum of a function by successively minimising along the coordinate directions.

1. *Find initial estimates of A_k^0 and B_k^0, which are denoted by*

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

   respectively. These initial estimates can be obtained by using the conventional (non-penalised) 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

$$\begin{aligned}
&\hat{\boldsymbol{\alpha}}_{-l}^{(j)}(U_s) = \left[\hat{\alpha}_1^{(j+1)}(U_s), \cdots, \hat{\alpha}_{l-1}^{(j+1)}(U_s), 0, \hat{\alpha}_{l+1}^{(j)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}^{(j)}(U_s)\right]^{\intercal}, \\
&\hat{\boldsymbol{\alpha}}'^{(j)}(U_s) = \left[\hat{\alpha}_1'^{(j)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}'^{(j)}(U_s)\right]^{\intercal}, \\
&\hat{Y}_{t,-l}^{(j)} = Y_t - \tilde{\mathbf{X}}_t \hat{\boldsymbol{\alpha}}_{-l}^{(j)}(U_s) - \tilde{\mathbf{X}}_t \hat{\boldsymbol{\alpha}}'^{(j)}(U_s)(U_t - U_s), \\
&\tilde{\mathbf{E}}_{nl} = \left(\tilde{E}_{nl,1}, \cdots, \tilde{E}_{nl,n}\right)^{\intercal}, \ \ \tilde{E}_{nl,s} = \frac{2}{nh} \sum_{t=1}^{n} \tilde{X}_{t,l|\tilde{K}} \hat{Y}_{t,-l}^{(j)} K_h(U_t, U_s).
\end{aligned}$$

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

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

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*

$$\begin{aligned}
&\hat{\boldsymbol{\alpha}}^{(j+1)}(U_s) = \left[\hat{\alpha}_1^{(j+1)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}^{(j+1)}(U_s)\right]^{\intercal}, \\
&\hat{\boldsymbol{\alpha}}_{-l}'^{(j)}(U_s) = \left[\hat{\alpha}_1'^{(j+1)}(U_s), \cdots, \hat{\alpha}_{l-1}'^{(j+1)}(U_s), 0, \hat{\alpha}_{l+1}'^{(j)}(U_s), \cdots, \hat{\alpha}_{\tilde{K}}'^{(j)}(U_s)\right]^{\intercal}, \\
&\check{Y}_{t,-l}^{(j)} = Y_t - \tilde{\mathbf{X}}_t \hat{\boldsymbol{\alpha}}^{(j+1)}(U_s) - \tilde{\mathbf{X}}_t \hat{\boldsymbol{\alpha}}_{-l}'^{(j)}(U_s)(U_t - U_s), \\
&\check{\mathbf{E}}_{nl} = \left(\check{E}_{nl,1}, \cdots, \check{E}_{nl,n}\right)^{\intercal}, \ \ \check{E}_{nl,s} = \frac{2}{nh} \sum_{t=1}^{n} \tilde{X}_{t,l|\tilde{K}} \check{Y}_{t,-l}^{(j)} [(U_t - U_s)/h] K_h(U_t, U_s).
\end{aligned}$$

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

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

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 convergence of the estimates is achieved.*

Our numerical studies in Sections 5 and 6 below show that the above iterative procedure has reasonably good finite-sample performance.

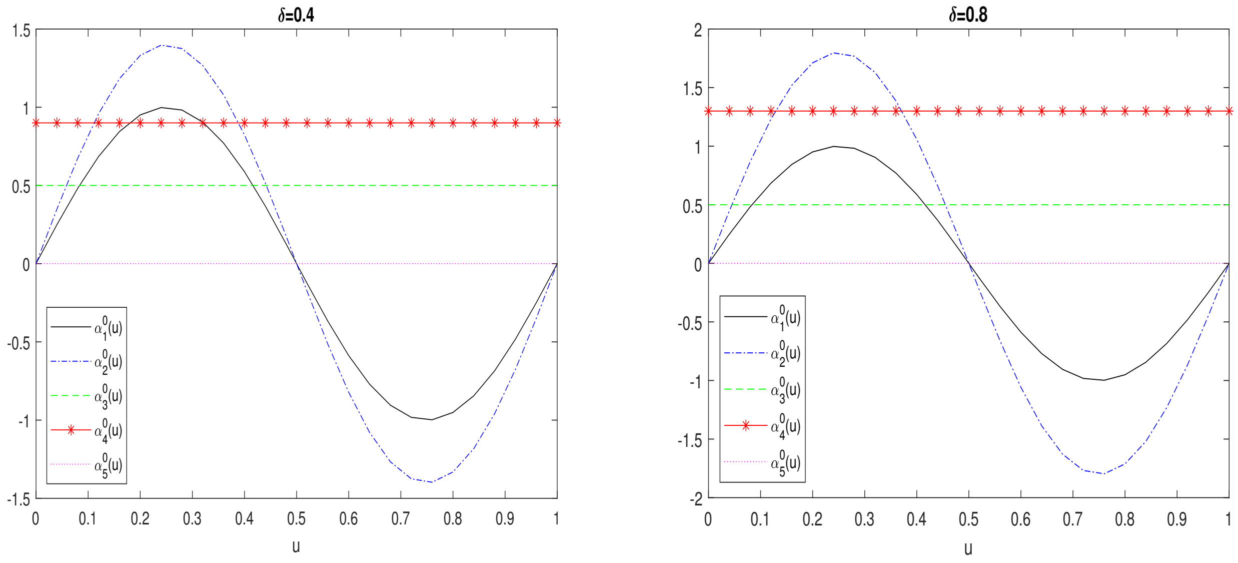


Figure 1: Plots of the cluster-specific coefficient functions. Left panel: $\delta = 0.4$; right panel: $\delta = 0.8$.

5 Monte-Carlo simulation

In this section, we conduct Monte-Carlo simulation studies to evaluate the finite-sample performance of the proposed methods.

Example 5.1. Consider the following functional-coefficient model:

$$Y_t = \sum_{j=1}^{p} \beta_j^0(U_t)X_{tj} + \sigma\varepsilon_t, \ \ t = 1, \cdots, n, \tag{5.1}$$

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 coefficient 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 (5.1) is defined as follows: $\beta^0_{4(k-1)+j}(\cdot) = \alpha^0_k(\cdot)$ for $k = 1, 2$ and $j = 1, 2, 3, 4$, and $\beta^0_{4(k-1)+j}(\cdot) = \alpha^0_k(\cdot) \equiv \alpha^0_k$ for $k = 3, 4, 5$ and $j = 1, 2, 3, 4$, where $\alpha^0_1(u) = \sin(2\pi u)$, $\alpha^0_2(u) = (1 + \delta)\sin(2\pi u)$, $\alpha^0_3 = 0.5$, $\alpha^0_4 = 0.5 + \delta$, $\alpha^0_5 = 0$, $\delta = 0.4$ or 0.8. The above homogeneity structure shows that there are five distinct clusters among the coefficients, some with functional coefficients varying with U_t and some with constant coefficients. The size of each cluster is the same. Figure 1 depicts these five coefficient functions.

The sample size n is set to be 200, 400 or 600, and the number of replications is $N = 500$. We first use the kernel method to obtain preliminary nonparametric estimates of the functional coefficients

$\beta_j^0(\cdot), j = 1, \cdots, 20$, with the Epanechnikov kernel $K(z) = \frac{3}{4}(1-z^2)_+$ and the optimal bandwidth selected from the cross-validation method in Section 4.1. The homogeneity and semi-varying coefficient structure in model (5.1) is ignored in this preliminary estimation. A combination of the kernel-based clustering method in Section 2.1 and the generalised information criterion in Section 2.2 is then used to estimate the latent homogeneity structure. In order to evaluate the clustering performance, we consider two commonly used measures: Normalised Mutual Information (NMI) and Purity, both of which can be used to examine how close the estimated set of clusters is 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 between $\mathcal{C}_1$ and $\mathcal{C}_2$ is defined as

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

where $H(\mathcal{C}_1)$ and $H(\mathcal{C}_2)$ are the entropies of $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, and $I(\mathcal{C}_1, \mathcal{C}_2)$ is the mutual information between $\mathcal{C}_1$ and $\mathcal{C}_2$ defined as:

$$I(\mathcal{C}_1, \mathcal{C}_2) = \sum_{k=1}^{K_1}\sum_{j=1}^{K_2}\left(\frac{|\mathcal{C}_k^1 \cap \mathcal{C}_j^2|}{p}\right)\log_2\left(\frac{p|\mathcal{C}_k^1 \cap \mathcal{C}_j^2|}{|\mathcal{C}_k^1||\mathcal{C}_j^2|}\right).$$

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

$$\mathsf{Purity}(\mathcal{C}_1, \mathcal{C}_2) = \frac{1}{p}\sum_{k=1}^{K_1}\max_{1\leqslant j\leqslant K_2}|\mathcal{C}_k^1 \cap \mathcal{C}_j^2|. \tag{5.2}$$

It is easy to find that the Purity measure also takes values between 0 and 1, and if $\mathcal{C}_1$ and $\mathcal{C}_2$ are equal, then $\mathsf{Purity}(\mathcal{C}_1, \mathcal{C}_2) = 1$. However, the purity measure does not trade off the quality of clustering against the number of clusters. For example, a purity value of 1 is achieved if one set contains singleton clusters. The NMI, by contrast, allows for this tradeoff.

Table 1 below presents the frequency (over 500 replications) at which a number between 1-10 is selected as the number of clusters by the information criterion detailed in Section 2.2. Table 2 gives the average values and standard errors (in parentheses) of the NMI and Purity measurements over 500 replications. From Table 1, we can find that when the covariates are uncorrelated, the number of clusters can be correctly estimated in about 80% of the replications even when $n = 200$ and $\delta = 0.4$, and when δ increases to 0.8, this percentage increases to almost 98%. When sample size increases to 400, the information criterion selects the correct number of clusters in almost all replications with uncorrelated covariates. When the correlation coefficient between the covariates is 0.25, the number of clusters is correctly estimated in only 34% of replications when $n = 200$ and

$\delta = 0.4$ and in over 70% of replications when $\delta = 0.8$. As the sample size increases to 400, this percentage rises to over 98%. In all of the specifications, the estimated number of clusters rarely goes below 3 or above 7. Table 2 shows that when there is no correlation among the 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 coefficient ρ from 0 to 0.25 causes the clustering precision to decrease. However, even for the case of $\delta = 0.4$ and $\rho = 0.25$, the NMI and purity values are close to 1 when the sample size is 400.

Table 1: Results on estimation of cluster number for Example 5.1

δ	ρ	n	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
0.4	0	200	0	0	3	54	400	39	4	0	0	0
		400	0	0	0	0	500	0	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0
0.4	0.25	200	0	0	146	157	170	25	2	0	0	0
		400	0	0	0	3	494	3	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0
0.8	0	200	0	0	0	1	489	9	1	0	0	0
		400	0	0	0	0	500	0	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0
0.8	0.25	200	0	0	15	62	365	45	12	1	0	0
		400	0	0	0	0	500	0	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0

Table 2: Average NMI and Purity and their standard deviations (in parentheses) for Example 5.1

		$\delta = 0.4$		$\delta = 0.8$	
ρ	n	NMI	Purity	NMI	Purity
0	200	0.9593 (0.0566)	0.9743 (0.0472)	0.9952 (0.0230)	0.9958 (0.0211)
	400	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	600	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
0.25	200	0.8459 (0.0880)	0.9503 (0.0567)	0.9368 (0.0774)	0.9596 (0.0583)
	400	0.9971 (0.0148)	0.9981 (0.0106)	1.0000 (0.0000)	1.0000 (0.0000)
	600	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)

We finally identify the clusters with zero coefficients and non-zero constant coefficients by using the penalised method introduced in Section 2.3. The tuning parameters in the penalty term are chosen by the GIC detailed in Section 4.1. In order to measure the accuracy of estimates of the functional coefficients $\beta_j^0(\cdot)$, $1 \leqslant j \leqslant p$, we compute the Mean Absolute Estimation Error (MAEE), which, for the preliminary (pre-clustering) kernel estimates, $\tilde{\beta}_j(\cdot)$, $1 \leqslant j \leqslant p$, is defined as

$$\mathsf{MAEE}(\mathsf{PreC} - \mathsf{Kernel}) = \frac{1}{np} \sum_{t=1}^{n} \sum_{j=1}^{p} |\tilde{\beta}_j(U_t) - \beta_j^0(U_t)|,$$

and for the post-clustering kernel estimates,

$$\mathsf{MAEE}(\mathsf{PostC} - \mathsf{Kernel}) = \frac{1}{np} \sum_{t=1}^{n} \sum_{j=1}^{p} |\tilde{\beta}_j^*(U_t) - \beta_j^0(U_t)|,$$

where $\tilde{\beta}_j^*(\cdot) = \tilde{\alpha}_k(\cdot)$ if $j \in \tilde{\mathcal{C}}_{k|\tilde{K}}$, $1 \leqslant k \leqslant \tilde{K}$, and $\tilde{\alpha}_k(\cdot) = \tilde{\alpha}_{k|\tilde{K}}(\cdot)$, $1 \leqslant k \leqslant \tilde{K}$, are the post-clustering kernel estimates of cluster-specific functional coefficients defined as in (2.4). Let $\hat{\beta}_j(\cdot) = \hat{\alpha}_k(\cdot)$ if $j \in \tilde{\mathcal{C}}_{k|\tilde{K}}$, $1 \leqslant k \leqslant \tilde{K}$, where $\hat{\alpha}_k(\cdot)$, $1 \leqslant k \leqslant \tilde{K}$, are the penalised estimates of the cluster-specific functional coefficients obtained by minimising (2.7). The MAEE of the penalised estimates is defined as

$$\mathsf{MAEE}(\mathsf{Penalised}) = \frac{1}{np} \sum_{t=1}^{n} \sum_{j=1}^{p} |\hat{\beta}_j(U_t) - \beta_j^0(U_t)|.$$

The main purpose for considering the MAEE of the post-clustering kernel and penalised estimates for $\beta_j^0(\cdot)$, $1 \leqslant j \leqslant p$, rather than for $\alpha_k^0(\cdot)$, $1 \leqslant k \leqslant K_0$, is to avoid having to order the estimated clusters and to match each of them to one of the true clusters (as there is no natural way to do this).

Table 3 below reports the average MAEE's and standard deviations (in parentheses) over 500 replications for the pre-clustering kernel estimation, post-clustering kernel estimation and the semiparametric penalised estimation. The results in the table show that, after identifying the homogeneity and semi-varying coefficient structure, the average MAEE values of the semiparametric penalised estimation are smaller than those of the post-clustering kernel estimation, which in turn are much smaller than those of the pre-clustering kernel estimation. In addition, all three estimation methods improve (with decreasing average MAEE values) as the sample size increases, and their performance becomes slightly worse when the correlation between the random covariates increases from 0 to 0.25.

Table 3: Average MAEE and standard deviation (in parentheses) for Example 5.1

δ	ρ	n	PreC-Kernel	PostC-Kernel	Penalised
0.4	0	200	0.1661 (0.0140)	0.0777 (0.0187)	0.0539 (0.0201)
		400	0.0967 (0.0056)	0.0447 (0.0035)	0.0260 (0.0058)
		600	0.0753 (0.0040)	0.0365 (0.0029)	0.0225 (0.0056)
0.4	0.25	200	0.2605 (0.0442)	0.1357 (0.0333)	0.1028 (0.0424)
		400	0.1441 (0.0097)	0.0560 (0.0060)	0.0253(0.0064)
		600	0.1090 (0.0055)	0.0445 (0.0034)	0.0200 (0.0042)
0.8	0	200	0.1918 (0.0161)	0.0778 (0.0132)	0.0460 (0.0132)
		400	0.1083 (0.0062)	0.0488 (0.0041)	0.0253 (0.0048)
		600	0.0832 (0.0043)	0.0393 (0.0029)	0.0223 (0.0037)
0.8	0.25	200	0.3020 (0.0522)	0.1336 (0.0439)	0.0845 (0.0541)
		400	0.1637 (0.0105)	0.0611 (0.0050)	0.0267 (0.0055)
		600	0.1206 (0.0054)	0.0492 (0.0037)	0.0233 (0.0048)

Example 5.2. We still consider model (5.1) with the following homogeneity structure: $\beta_1^0(\cdot) = \alpha_1^0(\cdot)$, $\beta_j^0(\cdot) = \alpha_2^0(\cdot)$ for $j = 2, 3$, $\beta_j^0(\cdot) \equiv \alpha_3^0$ for $j = 4, \cdots, 7$, $\beta_j^0(\cdot) \equiv \alpha_4^0$ for $j = 8, \cdots, 13$, and $\beta_j^0(\cdot) \equiv \alpha_5^0$ for $j = 14, \cdots, 20$. 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 5.1. The definitions of $\alpha_i^0(\cdot)$ and α_i^0 are also the same as those in the previous example. The sizes of clusters vary in this example, which are 1, 2, 4, 6, 7, respectively, while in Example 5.1 all clusters have an equal size of 4.

Tables 4 and 5 report the results for the estimation of the homogeneity structure and Table 6 reports the average MAEE values and standard deviations (in parentheses) for the pre-clustering kernel estimation, post-clustering kernel estimation and the penalised estimation over 500 replications. Although the sizes of clusters vary in this example, the results on the estimation of the number of clusters are similar to those in Example 5.1. The results on the clustering and estimation accuracy (especially the estimation accuracy) are better than those in Example 5.1, which is due to the fact that more coefficient functions (17 out of 20) take constant values in this example.

Table 4: Results on estimation of cluster number for Example 5.2

δ	ρ	n	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$
0.4	0	200	0	0	0	193	274	30	2	1	0	0
		400	0	0	0	4	495	1	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0
0.4	0.25	200	0	0	0	306	177	16	1	0	0	0
		400	0	0	0	43	457	0	0	0	0	0
		600	0	0	0	3	497	0	0	0	0	0
0.8	0	200	0	0	0	2	485	11	2	0	0	0
		400	0	0	0	0	499	1	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0
0.8	0.25	200	0	0	0	16	455	29	0	0	0	0
		400	0	0	0	0	500	0	0	0	0	0
		600	0	0	0	0	500	0	0	0	0	0

Table 5: Average NMI and Purity and their standard deviations (in parentheses) for Example 5.2

		$\delta = 0.4$		$\delta = 0.8$	
ρ	n	NMI	Purity	NMI	Purity
0	200	0.9785 (0.0308)	0.9901 (0.0365)	0.9979 (0.0120)	0.9973 (0.0178)
	400	0.9997 (0.0033)	0.9999 (0.0022)	0.9999 (0.0031)	0.9998 (0.0045)
	600	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
0.25	200	0.9630 (0.0420)	0.9869 (0.0379)	0.9900 (0.0302)	0.9918 (0.0280)
	400	0.9970 (0.0097)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
	600	0.9998 (0.0027)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)

Table 6: Average MAEE and standard deviation (in parentheses) for Example 5.2

δ	ρ	n	PreC-Kernel	PostC-Kernel	Penalised
0.4	0	200	0.1179 (0.0100)	0.0534 (0.0102)	0.0410 (0.0100)
		400	0.0755 (0.0049)	0.0358 (0.0042)	0.0167 (0.0040)
		600	0.0597 (0.0033)	0.0287 (0.0030)	0.0134 (0.0031)
0.4	0.25	200	0.1457 (0.0137)	0.0660 (0.0152)	0.0353 (0.0144)
		400	0.0919 (0.0059)	0.0383 (0.0051)	0.0173 (0.0054)
		600	0.0724 (0.0040)	0.0309 (0.0033)	0.0131 (0.0031)
0.8	0	200	0.1343 (0.0113)	0.0622 (0.0096)	0.0304 (0.0080)
		400	0.0843 (0.0050)	0.0394 (0.0042)	0.0188 (0.0042)
		600	0.0664 (0.0036)	0.0315 (0.0033)	0.0157 (0.0037)
0.8	0.25	200	0.1686 (0.0153)	0.0701 (0.0160)	0.0346 (0.0157)
		400	0.1030 (0.0066)	0.0414 (0.0046)	0.0203 (0.0093)
		600	0.0803 (0.0044)	0.0332 (0.0033)	0.0151 (0.0065)

6 Empirical applications

We next apply the developed model and methodology to the well-known Boston house price data. This data set has been previously analysed in many studies (c.f., Fan and Huang, 2005; Cai and Xu, 2008; Wang and Xia, 2009; Leng, 2010), where functional-coefficient models are usually recommended. However, it is not clear whether certain homogeneity structure among the functional coefficients exists. To investigate what factors influencing the house prices, we choose MEDV (the median value of owner-occupied homes in US \$1000) as the response variable and the following 13 variables as the explanatory variables: INT (the intercept), CHAS (Charles River dummy variable; =1 if tract bounds river, 0 otherwise), RAD (index of accessibility to radial highways), CRIM (crime rate per capita by town), ZN (proportion of residential land zoned for lots over 25000 sq. ft.), INDUS (proportion of non-retail business acres per town), NOX (nitric oxides concentration in parts per 10 million), RM (average number of rooms per dwelling), AGE (proportion of owner-occupied units built prior to 1940), DIS (weighted distances to five Boston employment centres), TAX (full-value property-tax rate per US \$10000), PTRATIO (pupil-teacher ratio by town), and B (1000(Bk-0.63)^2 where Bk is the proportion of blacks by town). The variable LSTAT (percentage of lower status population) is chosen as the index variable in the varying-coefficient model, which enables us to investigate the interaction of LSTAT with the explanatory

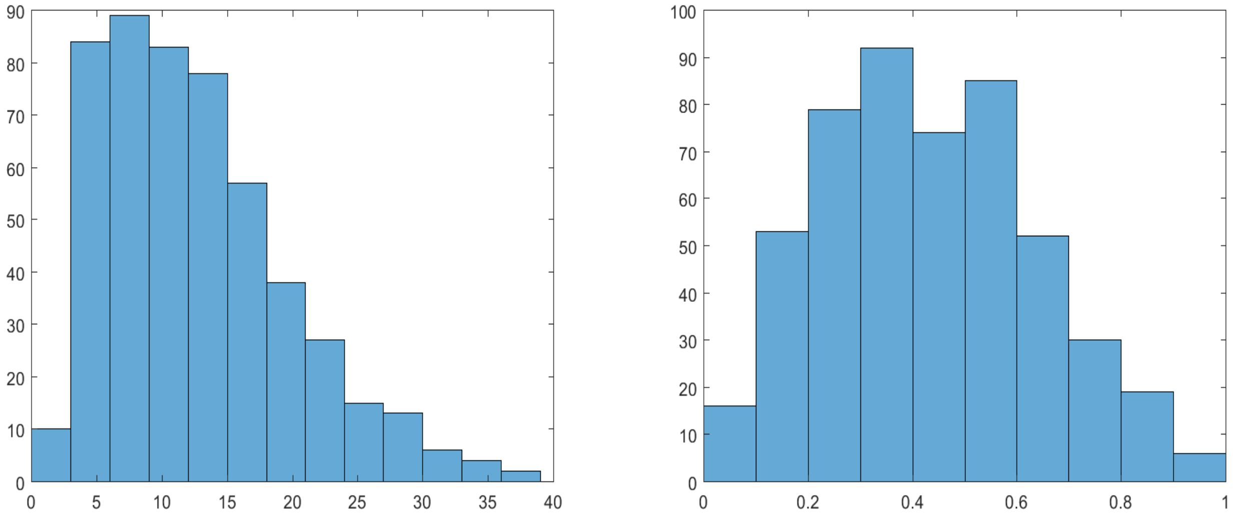


Figure 2: Histograms for the original and transformed index variable in Example 6.1. Left panel: original data for LSTAT; right panel: LSTAT after the square-root and min-max transformations.

variables. The sample size is $n = 506$. The response variable and all explanatory variables (except the intercept, INT) undergo the Z-score transformation before being fitted: i.e., for any variable, x_t, to be transformed, its Z-score is

$$z_t = \frac{x_t - \bar{x}}{s(x)}, \quad t = 1, \cdots, 506, \tag{6.1}$$

where $\bar{x}$ and $s(x)$ are the sample mean and sample standard deviation of x_t. Furthermore, as shown in the left panel of Figure 2, the index variable, LSTAT, exhibits strong skewness. Hence, we first take the square-root transformation of this variable to alleviate skewness and then the min-max normalisation:

$$U_t^\star = \frac{U_t - \min(U)}{\max(U) - \min(U)}, \tag{6.2}$$

where $\min(U)$ and $\max(U)$ denote the minimum and maximum of the observations of U, respectively. After the min-max normalisation, the support of $U_t^\star$ becomes $[0, 1]$, consistent with the assumption made on the index variable in the asymptotic theory. A histogram of this transformed variable is shown in the right panel of Figure 2.

Figure 3 plots the pre-clustering kernel estimated functional coefficients with the optimal bandwidth selected via the leave-one-out cross-validation method. The kernel-based clustering method and the generalised information criterion identify six clusters. The membership of these clusters and the characteristics of their functional coefficients are described in Table 7. DIS and TAX are found, by the penalised method, to have constant and similar negative effects on the response,

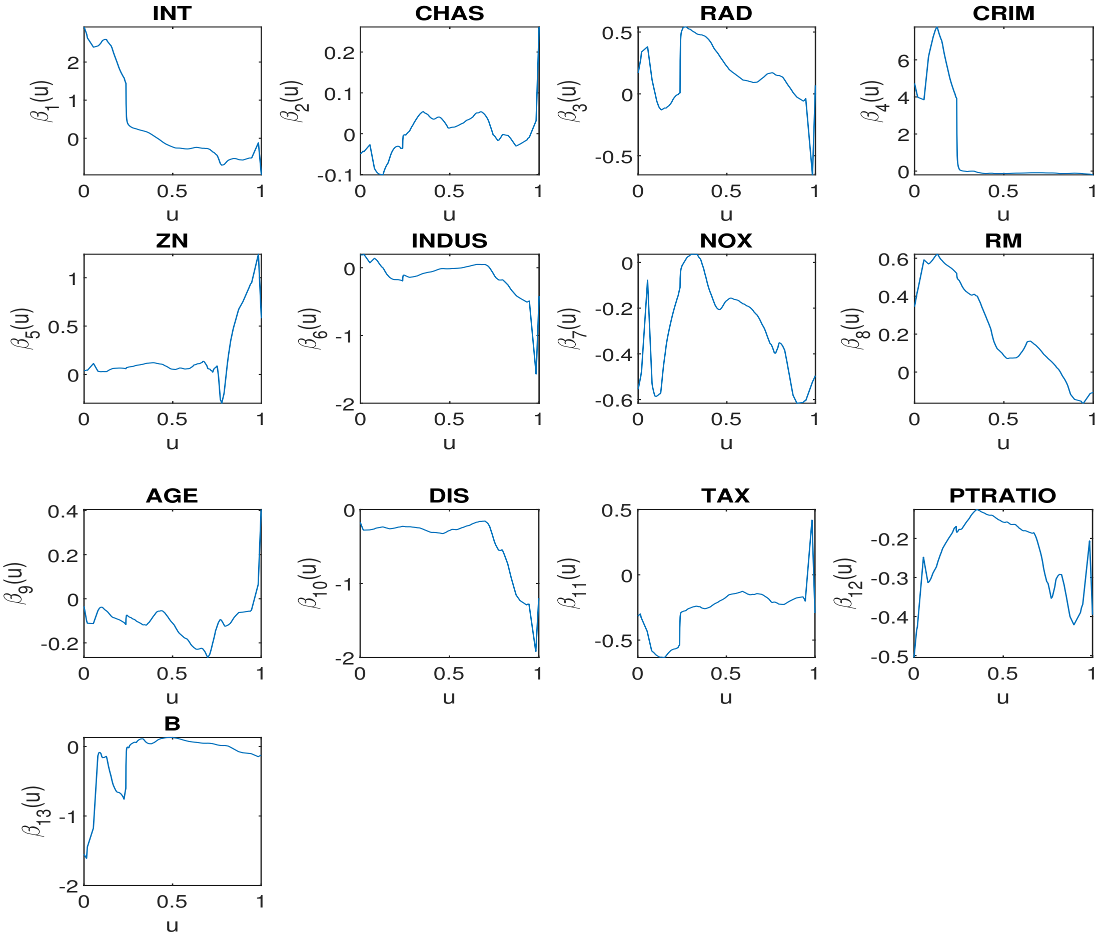


Figure 3: Pre-clustering estimates of the functional coefficients in Example 6.1.

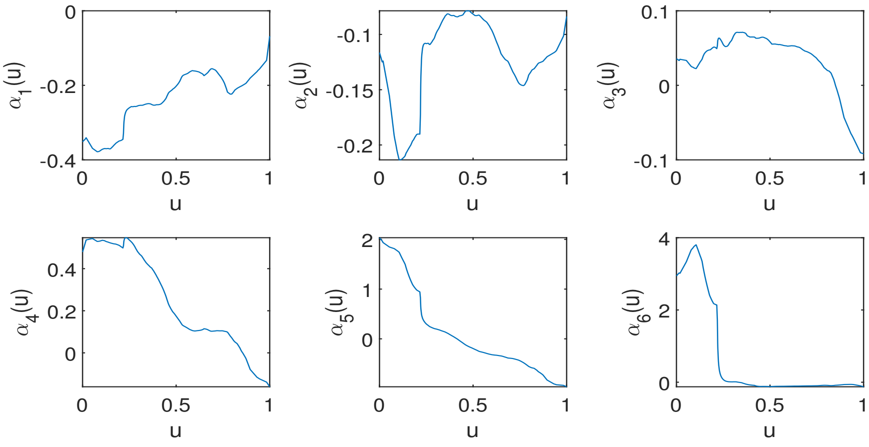


Figure 4: Post-clustering estimates of the functional coefficients in Example 6.1 with $\alpha_k(\cdot)$, for each $k = 1, 2, \ldots, 6$, being the estimated functional coefficient corresponding to the k-th cluster listed in Table 7.

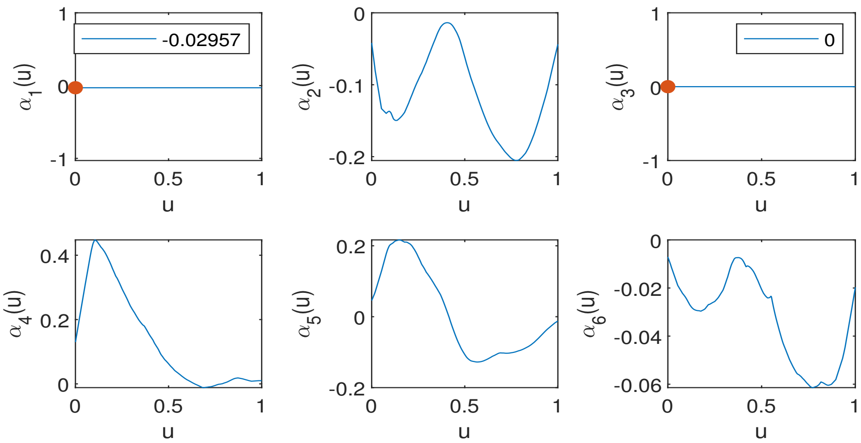


Figure 5: Penalised estimates of the functional coefficients in Example 6.1 with $\alpha_k(\cdot)$, for each $k = 1, 2, \ldots, 6$, being the estimated functional coefficient corresponding to the k-th cluster listed in Table 7.

while the variables, CHAS, ZN, and B are found to be insignificant. All the other explanatory variables have varying effects on the response as the value of LSTAT changes. Plots of the post-clustering kernel estimates of the functional coefficients and their penalised local linear estimates are shown in Figures 4 and 5, where for each $k = 1, \cdots, 6$, $\alpha_k(\cdot)$ denotes the functional coefficient corresponding to the k-th cluster listed in Table 7. The optimal tuning parameters in the penalised method are chosen, by the GIC, as $\lambda_1 = 10$ and $\lambda_2 = 2.3$.

Table 7: The estimated homogeneity structure in Example 6.1

Clusters	Variables	Coefficient functions
Cluster 1	DIS, TAX	Constant, value is -0.0296
Cluster 2	INDUS, NOX, AGE, PTRATIO	Non-constant, values are negative
Cluster 3	CHAS, ZN, B	Constant, value is 0
Cluster 4	RAD, RM	Non-constant, values are mostly positive
Cluster 5	INT	Non-constant
Cluster 6	CRIM	Non-constant, values are negative

We next compare the out-of-sample predictive performance between the pre-clustering (preliminary) kernel method, the post-clustering kernel method and the proposed penalised method. We randomly split the full sample into a training set of size 400 and a testing set of size 106 and repeat 200 times to reduce randomness in the results obtained. When calculating out-of-sample predictions for the post-clustering and penalised methods, we use the homogeneity structure (i.e.

the clusters and their membership) estimated from the full sample but estimate the values of the functional coefficients (evaluated at the LSTAT values belonging to the testing set) or the constant coefficients from the training sets. The predictive performance is measured by Mean Absolute Prediction Error (MAPE), which is defined by

$$\mathsf{MAPE} = \frac{1}{n_\star} \sum_{t=1}^{n_\star} \left| Y_t^\star - \hat{Y}_t^\star \right|, \tag{6.3}$$

where $n_\star$ is the size of the testing set (106 in this example), $Y_t^\star$ is a true value of the response variable in the testing sample, and $\hat{Y}_t^\star$ is the predicted value of $Y_t^\star$ using the model estimated from the training sample. Table 8 below reports the average MAPE values over 200 replications of random sample splitting. We consider bandwidth values in the range [0.06, 0.18] (with equal increment 0.02), which covers the optimal bandwidth of 0.168 for the preliminary kernel estimation and post-clustering kernel estimation. From Table 8, we can see that predicted values calculated from the model estimated by the penalised method have the smallest MAPE's over the range of bandwidth considered. Predictions made from the model estimated by the post-clustering kernel method have slightly larger MAPE values, while predictions from the pre-clustering kernel method has the largest MAPE values. This comparison result shows that the simplified functional-coefficient models from the developed kernel-based clustering and penalised methods provide more accurate out-of-sample prediction.

Table 8: Average MAPE over 200 times of random sample splitting in Example 6.1

Method	$h = 0.06$	$h = 0.08$	$h = 0.10$	$h = 0.12$	$h = 0.14$	$h = 0.16$	$h = 0.18$
PreC-Kernel	0.4957	0.4117	0.3622	0.3254	0.3029	0.2957	0.2944
PostC-Kernel	0.3436	0.3319	0.3091	0.2995	0.2946	0.2919	0.2919
Penalised	0.3273	0.3092	0.2987	0.2913	0.2858	0.2834	0.2844

7 Conclusion

In this paper, we have developed the kernel-based hierarchical clustering method and a generalised version of information criterion to uncover the latent homogeneity structure in the 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 Section 3 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 the empirical application to the Boston house price 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 absolute prediction errors using the developed approach are consistently smaller than those using the naive kernel method which ignores the latent homogeneity structure among the functional coefficients.

Supplementary materials

The online supplementary material contains the detailed proofs of Theorems 1-3.

References

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

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

Cai, Z. and Xu, X. (2008). Nonparametric quantile estimations for dynamic smooth coefficient models. *Jounal of the American Statistical Association*, **103**, 1595–1608.

Chen, J., Li, D. and Xia, Y. (2018). Estimation of a rank-reduced functional-coefficient panel data model in presence of serial correlation. *Working paper*.

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

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

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall, London.

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

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

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

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

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

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

Green, P. and Silverman, B. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC.

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

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

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

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

Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statistica Sinica*, **20**, 167–181.

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

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association*, **109**, 266–274.

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

Rencher, A. C. and Christensen, W. F. (2012). *Methods of Multivariate Analysis* (3rd Edition). Wiley Series in Probability and Statistics, Wiley.

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

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

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

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, **67**,91–108.

Vogt, M. and Linton, O. (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 Jones, M. C. (1994). *Kernel Smoothing*. Chapman and Hall.

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

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

Supplemental Document for “Nonparametric Homogeneity Pursuit in Functional-Coefficient Models”

Jia Chen[1], Degui Li[2], Lingling Wei[2], Wenyang Zhang[2]
[1]Department of Economics and Related Studies, University of York, United Kingdom
[2]Department of Mathematics, University of York, United Kingdom

9 February 2019

Appendix: Proofs of the asymptotic theorems

In this appendix, we give the detailed proofs of the main asymptotic results in Section 3 of the main document.

Proof of Theorem 1. From the definition of Δ_{ij}^0, we have $\Delta_{ij}^0 = 0$ if $i, j \in \mathcal{C}_k^0$; and $\Delta_{ij}^0 = \delta_{k_1 k_2}$ if $i \in \mathcal{C}_{k_1}^0$ and $j \in \mathcal{C}_{k_2}^0$ with $1 \leqslant k_1 \neq k_2 \leqslant K_0$, 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 (3.1), we only need to prove that

$$\max_{1 \leqslant i,j \leqslant p} \left|\tilde{\Delta}_{ij} - \Delta_{ij}^0\right| = o_P(\delta_n), \;\; \delta_n = \min_{1 \leqslant k_1 \neq k_2 \leqslant K_0} \delta_{k_1 k_2}. \tag{A.1}$$

From the definitions of $\tilde{\Delta}_{ij}$ and Δ_{ij}^0 in Section 2.1, it is sufficient to show

$$\max_{1 \leqslant i \leqslant p} \sup_{u \in \mathcal{U}_h} \left|\tilde{\beta}_i(u) - \beta_i^0(u)\right| = o_P(\delta_n). \tag{A.2}$$

In fact, if (A.2) holds, by the definition of $\tilde{\Delta}_{ij}$ and letting

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

we have

$$\max_{1 \leqslant i,j \leqslant p} \left|\tilde{\Delta}_{ij} - \Delta_{ij}\right| \leqslant 2 \max_{1 \leqslant i \leqslant p} \sup_{u \in \mathcal{U}_h} \left|\tilde{\beta}_i(u) - \beta_i^0(u)\right| = o_P(\delta_n). \tag{A.3}$$

For the case of $i, j \in \mathcal{C}_k^0$, we readily have $\Delta_{ij}^0 = \Delta_{ij} = 0$, and thus (A.3) leads to (A.1). On the other hand, uniformly for $i \in \mathcal{C}_{k_1}^0$ and $j \in \mathcal{C}_{k_2}^0$ with $1 \leqslant k_1 \neq k_2 \leqslant K_0$, as $n^{1/2}\delta_n/(\log n)^{1/2} \to \infty$ in Assumption 4(ii), we have

$$|\Delta_{ij} - \delta_{k_1 k_2}| = O_P\left(\sqrt{\log n / n}\right) = o_P(\delta_n), \tag{A.4}$$

which together with (A.3), implies that (A.1) holds.

We next prove (A.2). By (1.1) and (2.1), we have

$$\tilde{\boldsymbol{\beta}}(u)-\boldsymbol{\beta}_0(u) = \left[\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}K_h(U_t,u)\right]^{-1}\left[\sum_{t=1}^{n}\mathbf{X}_t\varepsilon_tK_h(U_t,u)\right]+ \left[\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}K_h(U_t,u)\right]^{-1}\left[\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^{\intercal}\boldsymbol{\beta}_t(u)K_h(U_t,u)\right], \tag{A.5}$$

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

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

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

$$\max_{1\leqslant i,j\leqslant p}\sup_{u\in\mathcal{U}_h}\left|\omega_{n,ij}(u)-\omega_{ij}^0(u)\right|=O_P\left(h^2+\epsilon_n\right), \tag{A.6}$$

where $\epsilon_n=\sqrt{\log h^{-1}/(nh)}$. Then, by (A.6) and Assumption 4(ii), we may show that

$$\sup_{u\in\mathcal{U}_h}\|\boldsymbol{\Omega}_n(u)-\boldsymbol{\Omega}_0(u)\|_F=O_P\left(p(\epsilon_n+h^2)\right)=o_P(1), \tag{A.7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. Using (A.7), Assumption 3(i) and Weyl's inequality, the smallest eigenvalue of $\boldsymbol{\Omega}_n(u)$ is positive and bounded away from zero uniformly for $u\in\mathcal{U}_h$, i.e.,

$$\inf_{u\in\mathcal{U}_h}\lambda_{\min}(\boldsymbol{\Omega}_n(u))>\zeta_0, \tag{A.8}$$

where ζ_0 is a positive constant.

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

$$\sup_{u\in\mathcal{U}_h}\left\|\frac{1}{nh}\sum_{t=1}^{n}\mathbf{X}_t\varepsilon_tK_h(U_t,u)\right\|=O_P\left(p^{1/2}\epsilon_n\right). \tag{A.9}$$

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

$$\sup_{u\in\mathcal{U}_h} \left\| \frac{1}{nh}\sum_{t=1}^{n} \mathbf{X}_t\mathbf{X}_t^\intercal \boldsymbol{\beta}_t(u) K_h(U_t, u) \right\| = O_P\left(p^{1/2}h^2\right). \tag{A.10}$$

Combining (A.5) and (A.8)–(A.10), we have

$$\sup_{u\in\mathcal{U}_h} \left\|\tilde{\boldsymbol{\beta}}(u) - \boldsymbol{\beta}_0(u)\right\| = O_P\left(p^{1/2}\epsilon_n + p^{1/2}h^2\right) = o_P(\delta_n), \tag{A.11}$$

which leads to (A.2). 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

$$\begin{aligned} \mathsf{P}\left(\tilde{K} = K_0\right) &= \mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n(K_0)\right) + \mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n^c(K_0)\right) \\ &= \mathsf{P}\left(\tilde{K} = K_0, \mathbf{C}_n(K_0)\right) + o(1). \end{aligned} \tag{A.12}$$

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

$$\mathsf{P}\left(\tilde{K} = K_0 | \mathbf{C}_n(K_0)\right) \to 1. \tag{A.13}$$

From the definition of $\tilde{K}$, (A.13) can be proved if the following result hold:

$$\mathsf{P}\left(\mathsf{IC}(K) > \mathsf{IC}(K_0),\ 1 \leqslant K \neq K_0 \leqslant \bar{K}\ | \mathbf{C}_n(K_0)\right) \to 1. \tag{A.14}$$

We consider (A.14) separately for the two cases: $1 \leqslant K \leqslant K_0 - 1$ and $K_0 + 1 \leqslant K \leqslant \bar{K}$. If $K_0 = 1$, the first case can be ignored. In fact, (A.15) in Proposition 1 below indicates that when $K_0 \leqslant K \leqslant \bar{K}$ and n is sufficiently large, $\mathsf{IC}(K)$ is a strictly increasing function of K, which proves (A.14) for the second case. On the other hand, for $1 \leqslant K \leqslant K_0 - 1$, 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 (A.14) for the first case. The proof of Theorem 2 has been completed. □

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

$$\mathsf{IC}(K) = \log(\sigma^2) + K \cdot \left[\frac{\log(nh)}{nh}\right]^{\rho} (1 + o_P(1)), \tag{A.15}$$

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

Proof. When $K_0 + 1 \leqslant K \leqslant \bar{K}$, 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 Section 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

$$\begin{aligned}
\tilde{\boldsymbol{\alpha}}_{K_0+1}(u_0) &= \left[\sum_{t=1}^n \mathbf{X}_{t,K_0+1}\mathbf{X}_{t,K_0+1}^{\intercal} K_h(U_t, u_0)\right]^{-1} \left[\sum_{t=1}^n \mathbf{X}_{t,K_0+1} Y_t K_h(U_t, u_0)\right] \\
&= \left[\frac{1}{nh}\sum_{t=1}^n \mathbf{X}_{t,K_0+1}\mathbf{X}_{t,K_0+1}^{\intercal} K_h(U_t, u_0)\right]^{-1} \left[\frac{1}{nh}\sum_{t=1}^n \mathbf{X}_{t,K_0+1}\varepsilon_t K_h(U_t, u_0)\right] + \\
&\quad \left[\frac{1}{nh}\sum_{t=1}^n \mathbf{X}_{t,K_0+1}\mathbf{X}_{t,K_0+1}^{\intercal} K_h(U_t, u_0)\right]^{-1} \left[\frac{1}{nh}\sum_{t=1}^n \mathbf{X}_{t,K_0+1}\mathbf{X}_{t,K_0+1}^{\intercal}\boldsymbol{\alpha}_{K_0+1}^*(U_t) K_h(U_t, u_0)\right] \\
&=: \boldsymbol{\Omega}_{n,K_0+1}^{-1}(u_0)\boldsymbol{\Lambda}_{n,K_0+1,\varepsilon}(u_0) + \boldsymbol{\Omega}_{n,K_0+1}^{-1}(u_0)\boldsymbol{\Lambda}_{n,K_0+1,\alpha}(u_0).
\end{aligned} \tag{A.16}$$

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

by Assumption 6(i), and furthermore,

$$\sup_{u\in\mathcal{U}_h}\|\boldsymbol{\Lambda}_{n,K_0+1,\varepsilon}(u)\|=O_P\left(p\epsilon_n\right),\ \sup_{u\in\mathcal{U}_h}\left\|\boldsymbol{\Lambda}_{n,K_0+1,\alpha}(u)-\boldsymbol{\alpha}^*_{K_0+1}(u)\right\|=O_P\left(p^2h^2\right), \tag{A.17}$$

where ϵ_n is defined in the proof of Theorem 1. Then, by (A.16) and (A.17), we have

$$\sup_{u\in\mathcal{U}_h}\left\|\tilde{\boldsymbol{\alpha}}_{K_0+1}(u)-\boldsymbol{\alpha}^*_{K_0+1}(u)\right\|=O_P\left(p\epsilon_n+p^2h^2\right)=O_P\left(p\epsilon_n\right) \tag{A.18}$$

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

Letting $I_t=I(U_t\in\mathcal{U}_h)$, 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

$$\begin{aligned}\tilde{\sigma}^2_n(K_0+1) &= \frac{1}{n_h}\sum_{t=1}^n\left[Y_t-\mathbf{X}^\intercal_{t,K_0+1}\tilde{\boldsymbol{\alpha}}_{K_0+1}(U_t)\right]^2I_t\\ &= \frac{1}{n_h}\sum_{t=1}^n\left[\varepsilon_t-\mathbf{X}^\intercal_{t,K_0+1}\left(\tilde{\boldsymbol{\alpha}}_{K_0+1}(U_t)-\boldsymbol{\alpha}^*_{K_0+1}(U_t)\right)\right]^2I_t\\ &= \frac{1}{n_h}\sum_{t=1}^n\varepsilon_t^2I_t+\frac{1}{n_h}\sum_{t=1}^n\varpi_t^2(K_0+1)I_t-\frac{2}{n_h}\sum_{t=1}^n\varepsilon_t\varpi_t(K_0+1)I_t, \end{aligned}\tag{A.19}$$

where $\varpi_t(K_0+1)=\mathbf{X}^\intercal_{t,K_0+1}\left(\tilde{\boldsymbol{\alpha}}_{K_0+1}(U_t)-\boldsymbol{\alpha}^*_{K_0+1}(U_t)\right)$. By some standard arguments and using (A.18), we may show that

$$\frac{1}{n_h}\sum_{t=1}^n\varepsilon_t^2I_t=\sigma^2+o_P(1), \tag{A.20}$$

$$\frac{1}{n_h}\sum_{t=1}^n\varpi_t^2(K_0+1)I_t=O_P\left(p^4\epsilon_n^2\right), \tag{A.21}$$

$$\begin{aligned}\frac{2}{n_h}\sum_{t=1}^n\varepsilon_t\varpi_t(K_0+1)I_t &= O_P\left(p^2(nh)^{-1}+p^2n^{-1}h^{-1/2}+p^2n^{-1/2}h^2\right)\\ &= o_P\left(p^4\epsilon_n^2\right),\end{aligned}\tag{A.22}$$

where the condition $nh^6=o(1)$ in Assumption 7 is used in proving (A.22). By (A.19)–(A.22), we have

$$
\begin{aligned}
\mathsf{IC}(K_0+1) &= \log\left[\tilde{\sigma}_n^2(K_0+1)\right] + (K_0+1)\cdot\left[\frac{\log(nh)}{nh}\right]^{\rho} \\
&= \log(\sigma^2) + (K_0+1)\cdot\left[\frac{\log(nh)}{nh}\right]^{\rho} + O_P(p^4\epsilon_n^2) \\
&= \log(\sigma^2) + (K_0+1)\cdot\left[\frac{\log(nh)}{nh}\right]^{\rho}(1+o_P(1)) \qquad \text{(A.23)}
\end{aligned}
$$

as $p = o\left([\log(nh)/(nh)]^{(\rho-1)/4}\right)$ in Assumption 7. Similarly, for any $K_0 \leqslant K \leqslant \bar{K}$ and n sufficiently large, we can also prove (A.15). Details are omitted here to save the space. □

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

$$
\mathsf{IC}(K) > \log(\sigma^2 + c_\alpha) + o_P(1) \qquad \text{(A.24)}
$$

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 (A.24) for the case of $1 \leqslant K \leqslant K_0 - 1$. 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

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

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

$$
\begin{aligned}
& Y_t - \tilde{\mathbf{X}}_{t,K_0-1}^{\intercal}\tilde{\boldsymbol{\alpha}}_{K_0-1}(U_t) \\
= \;& \varepsilon_t + \mathbf{X}_{t,K_0}^{\intercal}\left[\boldsymbol{\alpha}_0(U_t) - \boldsymbol{\alpha}_{K_0-1}^{\diamond}(U_t)\right] - \mathbf{X}_{t,K_0-1}^{\intercal}\left[\tilde{\boldsymbol{\alpha}}_{K_0-1}(U_t) - \boldsymbol{\alpha}_{K_0-1}(U_t)\right] \\
=: \;& \varepsilon_t + \varpi_{t1}(K_0-1) + \varpi_{t2}(K_0-1). \qquad \text{(A.25)}
\end{aligned}
$$

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 (A.25), we have

$$\begin{aligned}
&\sum_{t=1}^{n}\left[Y_t-\tilde{\mathbf{X}}_{t,K_0-1}^{\intercal}\tilde{\boldsymbol{\alpha}}_{K_0-1}(U_t)\right]^2 \mathsf{I}(U_t\in\mathcal{U}_h)\\
=\;&\sum_{t=1}^{n}\varepsilon_t^2 I_t+\sum_{t=1}^{n}\varpi_{t1}^2 I_t+\sum_{t=1}^{n}\varpi_{t2}^2 I_t+2\Big(\sum_{t=1}^{n}\varepsilon_t\varpi_{t1}I_t+\sum_{t=1}^{n}\varepsilon_t\varpi_{t2}I_t+\sum_{t=1}^{n}\varpi_{t1}\varpi_{t2}I_t\Big). \quad (A.26)
\end{aligned}$$

Using Assumption 5, we may show that

$$\frac{1}{n_h}\sum_{t=1}^{n}\varpi_{t1}^2 I_t>c_\alpha(1+o_P(1)). \quad (A.27)$$

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

$$\sum_{t=1}^{n}\varpi_{t2}^2 I_t=O_P\left(np^4\epsilon_n^2\right)=o_P(n),\;\;\sum_{t=1}^{n}\varepsilon_t\varpi_{t2}I_t=o_P\left(np^4\epsilon_n^2\right)=o_P(n). \quad (A.28)$$

Furthermore, we can also prove that

$$\sum_{t=1}^{n}\varepsilon_t\varpi_{t1}I_t=O_P\left(pn^{1/2}\right)=o_P(n), \quad (A.29)$$

$$\sum_{t=1}^{n}\varpi_{t1}\varpi_{t2}I_t=O_P\left(np^3\epsilon_n\right)=o_P(n) \quad (A.30)$$

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

Using (A.20) and (A.26)–(A.42), we readily have

$$\mathsf{IC}(K_0-1)>\log(\sigma^2+c_\alpha)+o_P(1). \quad (A.31)$$

Similarly, we can prove (A.31) for any $1\leqslant K\leqslant K_0-2$, 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

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

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

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*

$$\frac{1}{n}\left\|\widehat{\mathbf{A}}_n-\mathbf{A}_0\right\|^2=O_P\left(\frac{p^4}{nh}\right),\quad \frac{1}{n}\left\|\widehat{\mathbf{B}}_n-\mathbf{B}_0\right\|^2=O_P\left(\frac{p^4}{nh^3}\right) \tag{A.32}$$

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, Ke and Zhang (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

$$\mathcal{Q}_n\left(\mathbf{A}_0+\gamma_n\mathbf{U}_1,\mathbf{B}_0+\gamma_n\mathbf{U}_2/h\right)-\mathcal{Q}_n(\mathbf{A}_0,\mathbf{B}_0)=\mathcal{I}_n(1)+\mathcal{I}_n(2)+\mathcal{I}_n(3), \tag{A.33}$$

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

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

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 $\tilde{\mathbf{X}}_{t,K_0}=\mathbf{X}_{t,K_0}$,

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

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

$$\begin{aligned}\mathcal{I}_n(1) &= -\frac{2\gamma_n}{nh}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\varepsilon_t+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{d}_{ts}\right)\left[\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s}+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}(U_t-U_s)/h\right]K_h\left(U_t,U_s\right)\\ &\quad +\frac{\gamma_n^2}{nh}\sum_{s=1}^{n}\sum_{t=1}^{n}\left[\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s}+\mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{2s}(U_t-U_s)/h\right]^2 K_h\left(U_t,U_s\right)\\ &=: \mathcal{I}_n(4)+\mathcal{I}_n(5). \qquad \text{(A.34)}\end{aligned}$$

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

$$\begin{aligned}\mathcal{I}_n(5) &= \frac{\gamma_n^2}{nh}\sum_{s=1}^{n}(\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})\left[\sum_{t=1}^{n}\left(\mathbf{X}_{t,K_0}\mathbf{X}_{t,K_0}^{\intercal}\right)\otimes\mathbb{U}_{ts}K_h\left(U_t,U_s\right)\right](\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})^{\intercal}\\ &= \gamma_n^2\sum_{s=1}^{n}(\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})\left[\frac{1}{nh}\sum_{t=1}^{n}\left(\mathbf{X}_{t,K_0}\mathbf{X}_{t,K_0}^{\intercal}\right)\otimes\mathbb{U}_{ts}K_h\left(U_t,U_s\right)\right](\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})^{\intercal}\\ &= \gamma_n^2\sum_{s=1}^{n}(\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})\left[f_U(U_s)\boldsymbol{\Sigma}_{X|K_0}(U_s)\otimes\boldsymbol{\Sigma}_K+O_P\left(p^2h^2+p^2\epsilon_n\right)\right](\mathbf{u}_{1s}^{\intercal},\mathbf{u}_{2s}^{\intercal})^{\intercal}\\ &\geqslant \gamma_n^2(\zeta_1+o_P(1))\left(\|\mathbf{U}_1\|^2+\|\mathbf{U}_2\|^2\right), \qquad \text{(A.35)}\end{aligned}$$

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 $j \geqslant 1$. Observe that

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

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

$$\begin{aligned}
\mathsf{E}\left[\mathcal{I}_n^2(4,1)\right] &= \mathsf{E}\left[\left(\sum_{s=1}^{n}\sum_{t=1}^{n} \varepsilon_t \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} K_h\left(U_t, U_s\right)\right)^2\right] \\
&\leqslant n\sum_{s=1}^{n} \mathsf{E}\left[\left(\sum_{t=1}^{n} \varepsilon_t \mathbf{X}_{t,K_0}^{\intercal}\mathbf{u}_{1s} K_h\left(U_t, U_s\right)\right)^2\right] \\
&= \mathsf{O}\left(p^2 n^2 h\right) \cdot \|\mathbf{U}_1\|^2. \qquad \text{(A.36)}
\end{aligned}$$

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

$$\mathsf{E}\left[\mathcal{I}_n^2(4,2)\right] = \mathsf{O}\left(p^2 n^2 h\right) \cdot \|\mathbf{U}_2\|^2. \qquad \text{(A.37)}$$

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

$$\mathsf{E}\left[|\mathcal{I}_n(4,3)|\right] = \mathsf{O}\left(p^2 n^{3/2} h^3\right) \cdot \|\mathbf{U}_1\| \qquad \text{(A.38)}$$

and

$$\mathsf{E}\left[|\mathcal{I}_n(4,4)|\right] = \mathsf{O}\left(p^2 n^{3/2} h^3\right) \cdot \|\mathbf{U}_2\|. \qquad \text{(A.39)}$$

Following (A.36)–(A.39) and noting that $p^2 n h^5 = O(1)$ in Assumption 9, we may show that

$$\mathcal{I}_n(4) = O_P\left(\gamma_n^2 n^{1/2}\right) \cdot \left(\|\mathbf{U}_1\| + \|\mathbf{U}_2\|\right). \qquad \text{(A.40)}$$

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

$$\mathcal{I}_n(1) \geqslant \gamma_n^2(\zeta_1/2 + o_P(1))\left(\|\mathbf{U}_1\|^2 + \|\mathbf{U}_2\|^2\right). \qquad \text{(A.41)}$$

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

$$\begin{aligned}
\mathcal{I}_n(2) &= \sum_{k=1}^{K_0} p'_{\lambda_1}\left(\|\tilde{A}_k\|\right)\left(\|A_k^0+\gamma_n U_{1k}\|-\|A_k^0\|\right) \\
&\geqslant \sum_{k=1}^{K_0-1} p'_{\lambda_1}\left(\|\tilde{A}_k\|\right)\left(\|A_k^0+\gamma_n U_{1k}\|-\|A_k^0\|\right) \qquad \text{(A.42)}
\end{aligned}$$

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

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

which together with Assumption 8, indicates that

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

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 $\mathcal{I}_n(2) \geqslant 0$ with probability approaching one. Analogously, we can also show that $\mathcal{I}_n(3) \geqslant 0$ with probability approaching one. Hence, for any small $\epsilon > 0$ there exists sufficiently large $C > 0$ such that

$$\mathsf{P}\Big\{\inf_{(\mathbf{U}_1,\mathbf{U}_2)\in\mathcal{C}_n(C)} \mathcal{Q}_n\left(\mathbf{A}_0+\gamma_n\mathbf{U}_1, \mathbf{B}_0+\gamma_n\mathbf{U}_2/h\right) > \mathcal{Q}_n(\mathbf{A}_0,\mathbf{B}_0)\Big\} \geqslant 1-\epsilon \qquad \text{(A.43)}$$

for large n, which leads to (A.32), 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

$$\mathsf{P}\left(\|\widehat{A}_{K_0}\|=0\right) = \mathsf{P}\left(\|\widehat{A}_{K_0}\|=0|\tilde{\mathbf{C}}_n\right)\mathsf{P}\left(\tilde{\mathbf{C}}_n\right) + \mathsf{P}\left(\|\widehat{A}_{K_0}\|=0|\tilde{\mathbf{C}}_n^c\right)\mathsf{P}\left(\tilde{\mathbf{C}}_n^c\right), \qquad \text{(A.44)}$$

which together with Theorems 1 and 2, implies that

$$\mathsf{P}\left(\|\widehat{A}_{K_0}\|=0|\tilde{\mathbf{C}}_n\right) \to 1 \qquad \text{(A.45)}$$

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

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

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

$$\|\tilde{A}_{K_0}\| = \|A^0_{K_0}\| + O_P\left(n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n\right) = O_P\left(n^{1/2}p^2h^2 + n^{1/2}p\epsilon_n\right) = o_P(\lambda_1). \tag{A.46}$$

From the definition of $p'_{\lambda_1}(\cdot)$ and (A.46), when $\|\widehat{A}_{K_0}\| \neq 0$, we have

$$\|\mathcal{P}'_{n1}(A_{K_0})\| = \lambda_1 \tag{A.47}$$

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

$$\mathcal{L}'_{nk,1}(\widehat{\mathbf{A}}_n, \widehat{\mathbf{B}}_n) = \mathcal{P}'_{n1}(\widehat{A}_k) \tag{A.48}$$

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 (A.47), indicates that (A.48) 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

$$\mathsf{P}\left(\|\widehat{B}_k\| = 0,\ k = K_*, \cdots, K_0\right) \to 1. \tag{A.49}$$

The proof of Theorem 3 has been completed. □

References

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

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

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