UNIVERSITY of York

Discussion Papers in Economics

No. 18/15

Estimating Latent Group Structure in Time-Varying Coefficient Panel Data Models

Jia Chen

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

Estimating Latent Group Structure in Time-Varying Coefficient Panel Data Models

Jia Chen

Department of Economics and Related Studies, University of York, UK

Version: October 24, 2018

Abstract

This paper studies the estimation of latent group structures in heterogeneous time-varying coefficient panel data models. While allowing the coefficient functions to vary over cross sections provides a good way to model cross-sectional heterogeneity, it reduces the degree of freedom and leads to poor estimation accuracy when the time-series length is short. On the other hand, in a lot of empirical studies, it is not uncommon to find that heterogeneous coefficients exhibit group structures where coefficients belonging to the same group are similar or identical. This paper aims to provide an easy and straightforward approach for estimating the underlying latent groups. This approach is based on the hierarchical agglomerative clustering (HAC) of kernel estimates of the heterogeneous time-varying coefficients when the number of groups is known. We establish the consistency of this clustering method and also propose a generalised information criterion for estimating the number of groups when it is unknown. Simulation studies are carried out to examine the finite sample properties of the proposed clustering method as well as the post-clustering estimation of the group-specific time-varying coefficients. The simulation results show that our methods give comparable performance as the penalised-sieve-estimation based classifier Lasso approach by Su et al. (2018), but are computationally easier. An application to a cross-country growth study is also provided.

Keywords: Hierarchical agglomerative clustering; Generalised information criterion; Kernel estimation; Panel data; Time-varying coefficients.

1 Introduction

Analysis of panel data has become one of the most important areas in theoretical and applied econometrics. The double-index panel modelling framework facilitates the exploration of dynamic information over time span and heterogenous structure over cross sections. In the past few decades, there have been exciting developments in parametric and nonparametric panel model estimation and inference, see, for example, Arellano (2003), Su and Ullah (2011), Chen et al. (2012), Robinson (2012), Hsiao (2014) and the references therein. In the existing literature, it is typically assumed that the regression relationship between variables is invariant cross sectionally, leading to homogenous panel data models. However, such an assumption might be inappropriate in many practical applications when the data are collected from individuals with different characteristics or in different geographical locations. In the context of parametric linear panel data models, Ke et al. (2016) and Su et al. (2016) impose latent group structures on the constant regression coefficients, and respectively use the binary segmentation and shrinkage methods to detect and estimate the group structures. In this paper, we aim to study this problem in a more general setting by allowing the model regression coefficients to vary smoothly over time and the panel data to have general cross-sectional dependence.

Suppose that we have the panel observations: $(Y_{it}, \mathbf{X}_{it})$, $i = 1, \cdots, N$, $t = 1, \cdots, T$, which are allowed to be serially correlated over t and cross-sectionally dependent over i. The primary interest is to investigate the relationship between the response variable Y_{it} and the p-dimensional explanatory vector $\mathbf{X}_{it}$. Consider the following heterogenous time-varying coefficient panel data model:

$$Y_{it} = \alpha_i + \mathbf{X}_{it}'\boldsymbol{\beta}_{it} + \epsilon_{it}, \tag{1.1}$$

where α_i are individual specific effects, $\boldsymbol{\beta}_{it}$ are p-dimensional vectors of time-varying functional coefficients which are heterogeneous over i, and the model errors ϵ_{it} are stationary over time t but may be cross-sectionally dependent. As in Robinson (1989) and Cai (2007), we assume that $\boldsymbol{\beta}_{it}$ are smooth functions of scaled times:

$$\boldsymbol{\beta}_{it} = \boldsymbol{\beta}_i\big(\frac{t}{T}\big), \quad t = 1, \cdots, T, \quad i = 1, \cdots, N, \tag{1.2}$$

where $\boldsymbol{\beta}_i(\cdot)$ is a p-dimensional vector of functions satisfying some smoothness conditions. In model (1.1), we allow for the existence of heterogeneous intercept functions by letting the first element of $\mathbf{X}_{it}$ be one. With $\mathbf{X}_{it} = \big(1, X_{it,1}, \ldots, X_{it,p-1}\big)'$ and $\boldsymbol{\beta}_i(\frac{t}{T}) = \big(\beta_{i,0}(\frac{t}{T}), \beta_{i,1}(\frac{t}{T}), \ldots, \beta_{i,p-1}(\frac{t}{T})\big)'$, we can

rewrite equation (1.1) as

$$\begin{aligned} Y_{it} &= \alpha_i + \mathbf{X}_{it}'\boldsymbol{\beta}_i\big(\frac{t}{T}\big) + \epsilon_{it} \\ &= \alpha_i + \beta_{i,0}\big(\frac{t}{T}\big) + \sum_{k=1}^{p-1}\beta_{i,k}\big(\frac{t}{T}\big)X_{it,k} + \epsilon_{it}. \end{aligned} \tag{1.3}$$

As both α_i and $\beta_{i,0}(\cdot)$ appear in the intercept of the model, to disentangle α_i and $\beta_{i,0}(\cdot)$ from each other, we impose the identification condition $\sum_{t=1}^{T}\beta_{i,0}(t/T) = 0$ on the intercept functions (Boneva et al., 2015). An alternative is to assume $\sum_{i=1}^{N}\alpha_i = 0$ on the individual effects. In this paper we use $\sum_{t=1}^{T}\beta_{i,0}(t/T) = 0$ for convenience of estimation. This will become clearer in Section 2 when we develop the estimation procedure.

When the intercept functions $\beta_{i,0}(\cdot)$, $i = 1, \ldots, N$, are homogeneous, i.e., $\beta_{i,0}(\cdot) \equiv \beta_0(\cdot)$, equation (1.3) becomes a panel data model with a common time trend but heterogeneous time-varying slope coefficients. Further assuming homogeneity of time-varying slope coefficients, i.e., $\beta_{i,k}(\cdot) \equiv \beta_k(\cdot)$, $k = 1, \ldots, p-1$, gives the model considered in Li et al. (2011), of which the nonparametric trending panel model in Robinson (2012) is a special case. Panel data models with homogeneous time-varying coefficients have been extensively studied in the literature (to list a few, Li et al., 2011; Chen et al., 2012; Zhang et al., 2012; Chen and Huang, 2017), and their estimation and inference methods have been well developed.

Note that model (1.3) offers great flexibility for modelling cross-sectional heterogeneity and time-varying effects of regressors on the dependent variable. However, without considering any group structure for $\boldsymbol{\beta}_i(\cdot)$, we can only reply on the sample information from the i-th cross section to estimate the time-varying coefficient vector $\boldsymbol{\beta}_i(\cdot)$. This will lead to slow estimation convergence rates in large samples and unsatisfactory estimation accuracy in finite samples when the time series length T is not large enough. Consequently the benefits of panel data for giving a larger number of pooled observations cannot be reaped. On the other hand, in a lot of empirical studies using heterogeneous panel data models, researchers find group structures where coefficients within each group are homogeneous but heterogeneous across groups. Such group structures arise due to the similarity of some cross sections in certain characteristics such as their geographical location. Hence, in this paper we consider the case where there exists a latent group structure for the heterogenous time-varying coefficient functions, i.e., there exists a partition of the cross-sectional index set $\{1, 2, \cdots, N\}$, denoted by $\{\mathcal{G}_1, \cdots, \mathcal{G}_{K_0}\}$, such that

$$\boldsymbol{\beta}_i(\cdot) = \boldsymbol{\gamma}_k(\cdot) \quad \text{for} \quad i \in \mathcal{G}_k \quad \text{and} \quad \mathcal{G}_k \cap \mathcal{G}_j = \emptyset \quad \text{for} \quad k \neq j, \tag{1.4}$$

where $\emptyset$ denotes the empty set. We assume that the Lebesgue measure of $\big\{u \in [0,1] : \ \boldsymbol{\gamma}_k(u) \neq \boldsymbol{\gamma}_j(u), \ k \neq j\big\}$ is uniformly (over k and j) strictly larger than a positive constant, and the number of latent groups, K_0, is finite but may be unknown in practice. The aim of this paper is to uncover the latent

group structure (1.4) by estimating the number of groups K_0 and determining the membership of each index set $\mathcal{G}_k$, $k = 1, \cdots, K_0$. Consequently, a nonparametric estimation of the time-varying coefficient functions making use of the estimated group structure can be constructed, which has faster convergence rate than the naive nonparametric estimation ignoring the latent group structure.

Estimation of latent group structures in nonparametric panel data models has received increasing attention in recent years. Vogt and Linton (2017, 2018) introduce kernel-based clustering methods to estimate the latent structures of univariate regression functions in panel data. Su et al. (2018) consider the same model structure as (1.1), and use a sieve approximation for the time-varying coefficient functions and the so-called classifier LASSO method to estimate the latent structure. In this paper, we use a fundamentally different method and relax some restrictive model assumptions in Su et al. (2018) (say, the cross-sectional independence assumption). Partly motivated by Li et al. (2018), we combine the kernel estimation method of the heterogenous time-varying coefficient functions with the classic hierarchical agglomerative clustering (HAC) method to estimate the latent group structure. We then use a generalised information criterion to determine the unknown group number K_0. The advantages and novelty of our methods lie in the following aspects.

(i) When $\mathbf{X}_{it} \equiv 1$ and $\alpha_i \equiv 0$ for all i and t, our model becomes $Y_{it} = \beta_{i,0}(t/T) + \epsilon_{it}$, which is the model considered in Vogt and Linton (2018) with a fixed-design covariate. Vogt and Linton (2018) also use the classical HAC algorithm to cluster nonparametric regression curves but base the HAC on the complete linkage of a multi-scale distance statistic which maximises a normalised point-wise distance between two regression curves over a grid of bandwidth and covariate values. The multi-scale distance statistic is constructed using large-sample approximation of normalised point-wise distance maximised over a grid. Furthermore, although their method does not require the selection of a bandwidth, it does require the choice of a threshold parameter, π_{NT}, for estimating the number of groups. Applied to the fixed design model $Y_{it} = \beta_{i,0}(t/T) + \epsilon_{it}$, our method is more straightforward to implement. The second simulation example in Section 4.2 shows that, with a similar data generating process, our method performs at least as well as that of Vogt and Linton (2018). Our proposed method can be easily implemented in R or Matlab with readily available packages or functions for HAC algorithm. Although our method does require a selection of a smoothing parameter, i.e., the bandwidth for nonparametric estimation of the time-varying coefficient functions, such a selection problem has been extensively studied in the literature and one can easily use one of the existing methods, such as the leave-one-out cross validation to tackle it.

(ii) Su et al. (2018) first use the sieve estimation for the functional coefficients then use the classifier LASSO (C-LASSO) method, which is first introduced in Su et al. (2016), to simultaneously estimate the functional coefficient vectors and classify them into groups. This method does not have a closed form solution, and hence an iterative numerical method has to be used to obtain

an estimate of the latent groups. Hence, our method is implementationally easier. The first simulation study in Section 4.2 shows that our method works as well as that of Su et al. (2018).

The rest of the paper is organised as follows. In Section 2, we develop an easy-to-implement approach for estimating the latent group structure when the number of groups, K_0, is known and then propose an information criterion to estimate K_0 when it is unknown. Section 3 gives the consistency of the proposed clustering method and the method for estimating K_0. Section 4 provides 2 simulation examples, in which the data generating processes are similar to the simulation designs of Su et al. (2018) and Vogt and Linton (2018) to facilitate comparison of performance of our method against those of theirs. These are then followed by an empirical application to a cross-country economic growth study consisting of 100 countries across the globe. Section 5 concludes the paper. All the proofs are relegated to the appendix.

2 Estimation methodology

In this section, we first introduce a kernel-based HAC algorithm to estimate the latent groups by assuming that the number of groups, K_0, is known, and then propose a generalised information criterion to determine the number K_0.

2.1 Kernel based HAC algorithm

To illustrate the kernel-based clustering method for estimating the group structure, we first assume that the number of clusters, K_0 is pre-specified. The kernel-based clustering method applies the classic HAC algorithm to kernel estimates of the time-varying coefficients $\boldsymbol{\beta}_i(\cdot)$. To estimate $\boldsymbol{\beta}_i(\cdot)$, we first absorb α_i into $\beta_{i,0}(\cdot)$ and denote $\beta^*_{i,0}(\cdot) = \alpha_i + \beta_{i,0}(\cdot)$. Then model (1.3) can be written as

$$Y_{it} = \mathbf{X}'_{it}\boldsymbol{\beta}^*_i\big(\frac{t}{T}\big) + \epsilon_{it},$$

where $\boldsymbol{\beta}^*_i(t/T) = \big(\beta^*_{i,0}(t/T), \beta_{i,1}(t/T), \ldots, \beta_{i,p-1}(t/T)\big)'$. Assume that each coefficient function $\beta_{i,k}(\cdot)$, $i = 1, 2, \cdots, N$, $k = 0, 1, \cdots, p-1$ are continuous. For each $i = 1, \cdots, N$, and any $0 < u_0 < 0$, we may use the kernel smoothing method to estimate $\boldsymbol{\beta}^*_i(u_0)$:

$$\widehat{\boldsymbol{\beta}}^*_i(u_0) = \left[\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}'_{it} K\left(\frac{t - u_0 T}{Th}\right)\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{X}_{it} Y_{it} K\left(\frac{t - u_0 T}{Th}\right)\right], \tag{2.1}$$

where $\widehat{\boldsymbol{\beta}}_i^*(\cdot) = \big[\widehat{\beta}_{i,0}^*(\cdot), \widehat{\beta}_{i,1}(\cdot), \cdots, \widehat{\beta}_{i,p-1}(\cdot)\big]'$, $K(\cdot)$ is a kernel function and h is a bandwidth. From the definition of the above kernel estimation, it is easy to find that we only use the local sample information from the i-th cross section, so its finite-sample performance may be relatively poor when the time series length T is not large. We next extract the estimate of the individual intercept function from $\widehat{\beta}_{i,0}^*(\cdot)$. Denote

$$\widehat{Z}_{it} = Y_{it} - \sum_{k=1}^{p-1} \widehat{\beta}_{i,k}\big(\frac{t}{T}\big) X_{it,k}. \tag{2.2}$$

It is easy to see that

$$\widehat{Z}_{it} \approx \alpha_i + \beta_{i,0}\big(\frac{t}{T}\big) + \epsilon_{it}, \tag{2.3}$$

which implies

$$\frac{1}{T}\sum_{t=1}^{T} \widehat{Z}_{it} \approx \alpha_i + \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}, \tag{2.4}$$

given the identification condition $\sum_{t=1}^{T} \beta_{i,0}(t/T) = 0$. We can eliminate the individual effects α_i from (2.3) by subtracting equation (2.4) from it, i.e.,

$$\widehat{Z}_{it} - \frac{1}{T}\sum_{t=1}^{T} \widehat{Z}_{it} \approx \beta_{i,0}\big(\frac{t}{T}\big) + \epsilon_{it} - \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}.$$

Since $\frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} = O_P(1/\sqrt{T}) = o_P(1)$ when $T \to \infty$, we can estimate $\beta_{i,0}(\cdot)$ in the same way as in (2.1) but with $\mathbf{X}_{it}$ and Y_{it} replaced by 1 and $\widehat{Z}_{it}^c := \widehat{Z}_{it} - \frac{1}{T}\sum_{t=1}^{T} \widehat{Z}_{it}$, respectively. Denote the subsequent estimator by $\widehat{\beta}_{i,0}(\cdot)$ and combine it with the estimators of the slope coefficient functions above to form the estimator,

$$\widehat{\boldsymbol{\beta}}_i(\cdot) = \big[\widehat{\beta}_{i,0}(\cdot), \widehat{\beta}_{i,1}(\cdot), \cdots, \widehat{\beta}_{i,p-1}(\cdot)\big]',$$

of the original functional coefficient vector $\boldsymbol{\beta}_i(\cdot)$.

We next apply the classic HAC algorithm to the estimates of the individual functional coefficients constructed above to obtain an estimate of the latent groups. To this end, we first define a distance measure for the estimated coefficient function. For any $\widehat{\boldsymbol{\beta}}_i(\cdot)$ and $\widehat{\boldsymbol{\beta}}_j(\cdot)$, define a weighted L_q-distance between them as:

$$\widehat{\delta}_{ij} = \frac{1}{T}\sum_{t=1}^{T} \Big\|\widehat{\boldsymbol{\beta}}_i(t/T) - \widehat{\boldsymbol{\beta}}_j(t/T)\Big\|_q W(t/T), \tag{2.5}$$

where $\|\cdot\|_q$ denotes the L_q-norm for a vector, $q \geq 1$, and $W(\cdot)$ is a pre-specified non-negative weight function which trims out the scaled time points close to either 0 or 1, circumventing the well-known boundary effect in kernel estimation to unduly affecting the distance. Li et al. (2018) use the L_1-norm and choose $W(\cdot)$ as an indicator function to estimate the homogeneity structure

among the functional coefficients for independent cross-sectional data, whereas Vogt and Linton (2018) consider the L_∞-distance for classifying univariate regression functions. In the numerical studies in Section 4, we use the L_2-norm to measure the distance. Note that if the two indices i and j are from the same index set $\mathcal{G}_k$, we expect that the value of $\widehat{\delta}_{ij}$ will be small.

When the time span T tends to infinity, under some regularity conditions we may show that $\widehat{\boldsymbol{\beta}}_i(u)$ converges to the true functional coefficient vector $\boldsymbol{\beta}_i(u)$ uniformly over u and i, indicating that $\widehat{\delta}_{ij}$ defined in (2.5) would be a reasonable estimate of δ_{ij} defined as

$$\delta_{ij} = \int_0^1 \left\| \boldsymbol{\beta}_i(u) - \boldsymbol{\beta}_j(u) \right\|_q W(u) du. \tag{2.6}$$

Then, we let Δ_N be an $N \times N$ distance matrix with the (i,j)-th entry being δ_{ij}. Correspondingly, we let $\widehat{\Delta}_N$ be the estimated distance matrix of Δ_N with the (i,j)-th entry being $\widehat{\delta}_{ij}$. When $i = j$, it is easy to find that $\delta_{ij} = \widehat{\delta}_{ij} = 0$, indicating that the main diagonal elements of Δ_N and $\widehat{\Delta}_N$ are zero.

With the feasible distance matrix $\widehat{\Delta}_N$, we can apply the classic HAC method to explore the latent group structure among the individual functional coefficients. The HAC method has been commonly used in the past few decades, see, for example, Ward (1963), Hastie et al. (2009), Everitt et al. (2011) and the references therein. A recent extension to the kernel-based HAC method in nonparametric classification can be found in Li et al. (2018) and Vogt and Linton (2018). For the time being, we assume that K_0, the number of groups, is known a priori, and will later introduce an information criterion for estimating this number when it is unknown. We let $\widehat{\mathcal{G}}_1, \cdots, \widehat{\mathcal{G}}_{K_0}$ be the estimated index sets obtained via the following algorithm.

Step 1. Start with N groups with each individual unit forming a group.

Step 2. Search for the smallest off-diagonal element in $\widehat{\Delta}_N$ and merge the corresponding two groups. These two groups are closest to each other among all groups by the measure of distance used.

Step 3. Re-calculate the distances between the current groups and update the estimated distance matrix (with its size reduced after each merging). Here the distance between two groups $\mathcal{A}_1$ and $\mathcal{A}_2$ is defined as the furthest distance between any two estimated functional coefficient vectors with one from $\mathcal{A}_1$ and the other from $\mathcal{A}_2$.

Step 4. Repeat Steps 2 and 3 until the number of groups reaches K_0.

As with any clustering algorithm, in each iteration before the given number of groups is reached, we merge the two groups which have the smallest distance to each other among all groups. The measure of distance between groups impacts the clustering results. In this paper, we use the furthest distance (or the "complete linkage" in the clustering analysis literature) between members from two

groups to measure how far away they are from each other. Other possible distance measures are the closest distance (or "single linkage" in the clustering analysis literature) or the weighted average distance.

2.2 Selection of number of groups

The kernel-based HAC method above relies on prior information on the number of latent groups. However, this number is usually unknown in practical applications and needs to be determined via certain data-driven rule. Hence our next task is to develop such a rule. For a given value of K for the number of latent groups, we let $\widehat{\mathcal{G}}_{1|K}, \cdots, \widehat{\mathcal{G}}_{K|K}$ be the K estimated index sets from the kernel-based HAC method in Section 2.1. In this case, there are K different vectors of coefficient functions, denoted by $\boldsymbol{\gamma}_{1|K}(\cdot), \cdots, \boldsymbol{\gamma}_{K|K}(\cdot)$, to be estimated, and it is sensible to pool data from individual units belonging to the same estimated group in the kernel estimation. Specifically, with the estimated group structure we have the following time-varying coefficient panel model:

$$Y_{it} = \alpha_i + \gamma_{k|K,0}(\frac{t}{T}) + \sum_{j=1}^{p-1} \gamma_{k|K,j}(\frac{t}{T}) X_{it,j} + \epsilon_{it}, \quad i \in \widehat{\mathcal{G}}_{k|K}, \quad k = 1, \cdots, K, \tag{2.7}$$

whose group-specific coefficient functions $\boldsymbol{\gamma}_{k|K}(u_0) = \big[\gamma_{k|K,0}(u_0), \cdots, \gamma_{k|K,p-1}(u_0)\big]'$ can be estimated as

$$\widehat{\boldsymbol{\gamma}}_{k|K}(u_0) = \left[\sum_{i \in \widehat{\mathcal{G}}_{k|K}} \sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{it}' K\left(\frac{t - u_0 T}{Th}\right)\right]^{-1} \left[\sum_{i \in \widehat{\mathcal{G}}_{k|K}} \sum_{t=1}^{T} \mathbf{X}_{it} Y_{it}^c K\left(\frac{t - u_0 T}{Th}\right)\right] \tag{2.8}$$

for $k = 1, \cdots, K$, and any $u_0 \in (0, 1)$. In (2.8), we have used the notation

$$Y_{it}^c = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} \widehat{Z}_{it},$$

where $\widehat{Z}_{it}$ was defined in (2.2). Note that we use Y_{it}^c instead of Y_{it} in (2.8). This is mainly to eliminate the individual effects α_i that may cause estimation bias in the above pooled kernel method.

We then define the following information criterion:

$$\mathbb{IC}(K) = \log \mathbb{V}_n^2(K) + K \cdot \rho, \tag{2.9}$$

where ρ is a tuning parameter whose value may rely on N, T, and h (due to the nonparametric

kernel-based estimation of the time-varying coefficients in the panel model), and

$$\mathbb{V}_n^2(K) = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\widehat{\mathcal{G}}_{k|K}}\sum_{t=1}^{T}\left[Y_{it}^c - \mathbf{X}_{it}'\widehat{\boldsymbol{\gamma}}_{k|K}(t/T)\right]^2 W(t/T).$$

The number of latent groups can be estimated by minimising the criterion $\mathbb{IC}(K)$, i.e.,

$$\widehat{K} = \arg\min_{1\le K\le \widetilde{K}} \mathbb{IC}(K), \tag{2.10}$$

where $\widetilde{K}$ is a pre-specified upper bound for the number of latent groups.

In Section 3 below, we will show that the estimator $\widehat{K}$, defined in (2.10), is a consistent estimate of the true cluster number K_0. To achieve the consistency property, we need to impose some mild restriction on the tuning parameter ρ in the penalty term (see Appendix A). Section 4 will discuss the practical choice of ρ in numerical studies. In practical data analysis, one first obtains $\widehat{K}$ from (2.9) and (2.10), and then use the kernel-based HAC procedure in Section 2.1 to identify the group membership of $\mathcal{G}_k$ by stopping the algorithm when the number of groups reaches $\widehat{K}$.

3 Large-sample theory

In this section we establish the asymptotic property of the methodology proposed in Sections 2.1 and 2.2. Theorem 1 shows that the kernel-based HAC algorithm can consistently estimate the membership of the latent groups $\mathcal{G}_k$, $k = 1, \cdots, K_0$, when the number K_0 is known.

THEOREM 1. *Suppose that Assumptions 1–4 in Appendix A are satisfied. If K_0, the number of latent groups, is known a priori, then*

$$\mathbb{P}\left(\{\widehat{\mathcal{G}}_1, \cdots, \widehat{\mathcal{G}}_{K_0}\} = \{\mathcal{G}_1, \cdots, \mathcal{G}_{K_0}\}\right) \to 1 \tag{3.1}$$

as $T \to \infty$.

REMARK 1. The consistency result in Theorem 1 is similar to some results in existing literature (although in different model settings), such as Theorem 3.1 in Vogt and Linton (2017), Theorem 1 in Li et al. (2018) and Theorem 4.1 in Vogt and Linton (2018). Note that we only require that T tends to infinity in Theorem 1. So the above result is applicable to settings where the cross-sectional size is either fixed or divergent to infinity. In addition, it is worth mentioning that we allow arbitrary cross-sectional dependence in derivation of Theorem 1.

THEOREM 2. *Suppose that Assumptions 1–6 in Appendix A are satisfied. Then we have*

$$\mathbb{P}\left(\widehat{K} = K_0\right) \to 1. \tag{3.2}$$

as $T \to \infty$.

REMARK 2. Su et al. (2018) also propose an information criterion for selecting the number of groups for their C-Lasso based clustering method and establish a similar consistency result under $N, T \to \infty$. Vogt and Linton (2018) use a thresholding method to choose the number of groups, which is also shown to be consistent. Li et al. (2018) also establish the consistency of their information criterion for choosing the number of homogeneous groups among functional coefficients for independent cross-sectional data. We note that in Theorem 2 we allow for the existence of cross-sectional dependence that satisfies Assumption 6 (especially between cross sections belonging to the same group). Furthermore, as in Theorem 1, the consistency result (3.2) holds whether N is fixed or diverging to infinity at a slower rate than T^m, where m is a positive constant defined in Assumption 4.

4 Numerical studies

In this section, we first discuss how to choose the bandwidth h and the tuning parameter ρ in Section 4.1 and then provide two Monte-Carlo experiments in Section 4.2 to demonstrate the finite-sample performance of the proposed methodology for identifying latent groups. Finally in Section 4.3, we apply our method to a cross-country economic growth study and discover 4 groups of countries which have distinct growth patterns.

4.1 Choice of tuning parameters

To achieve good grouping results, it is desirable to first obtain accurate nonparametric estimates of the functional coefficients, which, in turn, requires a proper choice of the bandwidth h. As the aim is to achieve good estimation accuracy, we can use existing bandwidth selection methods such as the leave-one-out cross-validation. This method selects the h value which minimises the following mean squared error

$$\mathbb{CV}(h) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[Y_{it}^{c} - \mathbf{X}_{it}'\widehat{\boldsymbol{\beta}}_{i,h}^{(-t)}(\frac{t}{T})\right]^2,$$

where, for each $i = 1, \ldots, N$ and $t = 1, \ldots, T$, Y_{it}^{c} was defined in Section 2.2 (which relies on h as the construction of $\widehat{Z}_{it}$ involves the nonparametric kernel estimates of the coefficient functions), and $\widehat{\boldsymbol{\beta}}_{i,h}^{(-t)}(\frac{t}{T})$ is the nonparametric estimate (with bandwidth h) of $\boldsymbol{\beta}_i(\frac{t}{T})$ obtained by using observations from the i-th cross section except the t-th observation $(Y_{it}, \mathbf{X}_{it}, t/T)$. The simulation studies in

Section 4.2 below show that such a selected bandwidth gives accurate estimation of the functional coefficients and good clustering results.

A proper choice of the tuning parameter ρ is crucial in order for the information criterion to work well. In our numerical study, we choose ρ as

$$\rho_1 = \frac{\log(N_K Th)}{N_K Th} \quad \text{or} \quad \rho_2 = \frac{2}{N_K Th} \text{ with } N_K = \min\left\{\left|\widehat{\mathcal{G}}_{k|K}\right|, \quad k = 1, \cdots, K\right\}, \tag{4.1}$$

where $|\mathcal{A}|$ denotes the cardinality of a set $\mathcal{A}$. This corresponds to a generalised Bayesian information criterion (GBIC with $\rho = \rho_1$) or generalised Akaike information criterion (GAIC with $\rho = \rho_2$) by treating $N_K Th$ as effective sample size (for the smallest cluster when the number of clusters is K). Such a criterion for estimating the number of latent groups works well in our simulation studies in Section 4.2. A similar criterion can also be found in Wang and Xia (2009) and Li et al. (2018) for variable selection and structure identification in high-dimensional varying-coefficient models for independent cross-sectional data.

4.2 Simulation studies

For easier comparison with the methods in Su et al. (2018) and Vogt and Linton (2018), we adopt a data generating process, i.e. DGP 2, from Su et al. (2018) in the first simulation study and then the data generating process from Section 7 of Vogt and Linton (2018) but with a fixed-design covariate in accordance with our modelling framework.

Simulated Example 1. This data generating process is the same as DGP 2 in Su et al. (2018),

$$Y_{it} = \alpha_i + \beta_{i,0}(\frac{t}{T}) + \beta_{i,1}(\frac{t}{T})X_{it} + \epsilon_{it}, \quad i = 1, \ldots, N, \ t = 1 \ldots, T,$$

where α_i and ϵ_{it} are independently drawn from the $N(0,1)$ distribution and are mutually independent,

$$\beta_{i,0}(u) = \begin{cases} \gamma_{1,0}(u) = 3F(u; 0.5, 0.1) & \text{if } i \in \mathcal{G}_1, \\ \gamma_{2,0}(u) = 3[2u - 6u^2 + 4u^3 + F(u; 0.7, 0.05)] & \text{if } i \in \mathcal{G}_2, \\ \gamma_{3,0}(u) = 3[4u - 8u^2 + 4u^3 + F(u; 0.6, 0.05)] & \text{if } i \in \mathcal{G}_3, \end{cases} \tag{4.2}$$

$$\beta_{i,1}(u) = \begin{cases} \gamma_{1,1}(u) = 3\big[2u - 4u^2 + 2u^3 + F(u; 0.6, 0.1)\big] & \text{if } i \in \mathcal{G}_1, \\ \gamma_{2,1}(u) = 3\big[u - 3u^2 + 2u^3 + F(u; 0.7, 0.04)\big] & \text{if } i \in \mathcal{G}_2, \\ \gamma_{3,1}(u) = 3\big[0.5u - 0.5u^2 + F(u; 0.4, 0.07)\big] & \text{if } i \in \mathcal{G}_3, \end{cases} \tag{4.3}$$

in which $F(u; \mu, \nu) = \frac{1}{1+\exp[-(u-\mu)/\nu]}$, $\mathcal{G}_1 = \{1, 2, \ldots, N_1\}$, $\mathcal{G}_2 = \{N_1 + 1, N_1 + 2, \ldots, N_1 + N_2\}$, and $\mathcal{G}_3 = \{N_1 + N_2 + 1, N_1 + N_2 + 2, \ldots, N_1 + N_2 + N_3\}$, and the cardinalities of the three groups are defined as $N_1 = 0.3N$, $N_2 = 0.3N$ and $N_3 = 0.4N$. The intercept functional coefficients, $\beta_{i,0}(t/T)$,

are demeaned so as to satisfy the identification condition $\sum_{t=1}^{T} \beta_{i,0}(t/T) = 0$. Different sample sizes of $N = 50, 100$ and $T = 40, 80$ are considered, and for each combination of N and T, 200 replicate samples are drawn from the data generating process. The bandwidth used for the nonparametric estimation of $\beta_i(\cdot)$ is selected using the leave-one-out cross validation method detailed in Section 4.1, and the kernel function used is the Epanechnikov kernel $K(u) = 3(1-u^2)_+/4$, where $(v)_+ = \max\{v, 0\}$.

For each combination of N and T, we report the accuracy of both the clustering and the estimation of the time-varying coefficients. To measure clustering accuracy, we calculate the purity and normalised mutual information (NMI) of our estimated clusters $\widehat{\mathcal{C}} = \{\widehat{\mathcal{G}}_1, \ldots, \widehat{\mathcal{G}}_{\widehat{K}}\}$ with the true clusters $\mathcal{C}_0 = \{\mathcal{G}_1, \ldots, \mathcal{G}_{K_0}\}$, which are defined, respectively, as

$$\text{Purity}(\widehat{\mathcal{C}}, \mathcal{C}_0) = \frac{1}{N} \sum_{k=1}^{\widehat{K}} \max_{1 \le j \le K_0} \left| \widehat{\mathcal{G}}_k \cap \mathcal{G}_j \right|$$

and

$$\text{NMI}(\widehat{\mathcal{C}}, \mathcal{C}_0) = \frac{\text{I}(\widehat{\mathcal{C}}, \mathcal{C}_0)}{(H(\widehat{\mathcal{C}}) + H(\mathcal{C}_0))/2},$$

where $\text{I}(\widehat{\mathcal{C}}, \mathcal{C}_0)$ is the mutual information between $\widehat{\mathcal{C}}$ and $\mathcal{C}_0$ defined as

$$\text{I}(\widehat{\mathcal{C}}, \mathcal{C}_0) = \sum_{k=1}^{\widehat{K}} \sum_{j=1}^{K_0} \left(\frac{|\widehat{\mathcal{G}}_k \cap \mathcal{G}_j|}{N} \right) \log_2 \left(\frac{N|\widehat{\mathcal{G}}_k \cap \mathcal{G}_j|}{|\widehat{\mathcal{G}}_k||\mathcal{G}_j|} \right),$$

and $H(\widehat{\mathcal{C}})$ is the entropy of $\widehat{\mathcal{C}}$ defined as

$$\text{H}(\widehat{\mathcal{C}}) = -\sum_{k=1}^{\widehat{K}} \frac{|\widehat{\mathcal{G}}_k|}{N} \log_2 \left(\frac{|\widehat{\mathcal{G}}_k|}{N} \right)$$

and $H(\mathcal{C}_0)$ is defined similarly. The advantage of using the measures of NMI and purity is that the results do not depend on the ordering of clusters in $\widehat{\mathcal{C}}$ or $\mathcal{C}_0$. The closer the values of NMI and purity are to 1, the more accurate the estimated clusters are to the true clusters. To measure estimation accuracy, we calculate the root mean squared errors (RMSE) of three estimators of $\beta_i(\cdot)$: the oracle estimator (obtained by assuming the true group structure is known a priori and pooling data from members of each group to obtain group-specific estimates of the coefficient functions), the pre-clustering estimator (obtained individual by individual without considering the group structure), and the post-clustering estimator (obtained by pooling data from members of each estimated group

Table 4.1: Frequencies at which K_0 is estimated for Simulated Example 1

Sample size		GBIC					GAIC				
		1	2	3(true)	4	5	1	2	3(true)	4	5
$N = 50$	$T = 40$	0	13	181	6	0	0	5	182	13	0
	$T = 80$	0	0	200	0	0	0	0	199	1	0
$N = 100$	$T = 40$	0	8	191	1	0	0	4	182	12	2
	$T = 80$	0	0	200	0	0	0	0	200	0	0

for group-specific estimates). Here the RMSE of an estimator $\widehat{\boldsymbol{\beta}}(\cdot) = \big(\widehat{\boldsymbol{\beta}}_1(\cdot), \ldots, \widehat{\boldsymbol{\beta}}_N(\cdot)\big)'$ is defined as

$$\text{RMSE}(\widehat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N}\Big\{\frac{1}{T}\sum_{t=1}^{T}\big\|\widehat{\boldsymbol{\beta}}_i(\frac{t}{T}) - \boldsymbol{\beta}_i(\frac{t}{T})\big\|_2^2\Big\}^{1/2}, \tag{4.4}$$

where those $\widehat{\boldsymbol{\beta}}_i(\cdot)$'s belonging to the same group in the oracle or post-clustering estimation are equal.

We first report, in Table 4.1, the frequency at which a certain number of groups is chosen over 200 replications. Then in Table 4.2 we summarise the average and standard deviation (in parentheses) of the purities and NMI's between the kernel based HAC results and the true group structure over these 200 replications. The average and standard deviation (in parentheses) of the RMSE's for the oracle, pre-clustering, and post-clustering estimation of the $\beta_i(\cdot)$'s are presented in Table 4.3.

Table 4.1 shows that the GBIC chooses the correct number of groups in about 91% of the repeated samples when the time series length T is 40 and this percentage rises to almost 100 when T increases to 80, irrespective of whether $N = 50$ or 100. These results are comparable to those in DGP 1 of Su et al. (2018), which are obtained from an information criterion deduced from their C-Lasso method. The GAIC has very similar performance in all the four combinations of N and T, which subsequently leads to the GBIC and GAIC having similar NMI and purity values as well as post-clustering estimation accuracy (measured by RMSE), as demonstrated by Tables 4.2 and 4.3. The NMI value for both the GBIC and GAIC is between 0.83-0.85 when $T = 40$ and then rises to around 0.98 when $T = 80$, and the purity is between 0.93-0.94 when $T = 40$ and then rises to more than 0.99 when $T = 80$. The RMSE's of the GBIC and GAIC post-clustering estimation of the functional coefficients are close to those of the oracle estimation. They are 50%-60% of the RMSE's of the pre-clustering nonparametric kernel estimation, a 40%-50% reduction, which shows the benefit of pooling data from cross sections of the same group for estimation.

Simulated Example 2. This data generating process is the same as that in Section 7 of Vogt and Linton (2018), except that we now replace the i.i.d. Uniform $[0, 1]$ exogenous variable X_{it} with the

Table 4.2: The average(standard deviation) NMI's and purities for Simulated Example 1

Sample size		GBIC		GAIC	
		NMI	Purity	NMI	Purity
$N = 50$	$T = 40$	0.8473(0.0980)	0.9408(0.0570)	0.8465(0.0989)	0.9304(0.0672)
	$T = 80$	0.9772(0.0441)	0.9925(0.0161)	0.9770(0.0449)	0.9919(0.0205)
$N = 100$	$T = 40$	0.8474(0.0754)	0.9470(0.0389)	0.8467(0.0751)	0.9370(0.0603)
	$T = 80$	0.9822(0.0295)	0.9952(0.0087)	0.9822(0.0295)	0.9952(0.0087)

Table 4.3: The average(standard deviation) RMSE's of $\beta_i(\cdot)$ estimates for Simulated Example 1

Sample size		Oracle	Pre-clustering	Post-clustering	
				GBIC	GAIC
$N = 50$	$T = 40$	0.2508(0.0145)	0.4856(0.0156)	0.2932(0.0431)	0.2908(0.0393)
	$T = 80$	0.1917(0.0122)	0.3618(0.0118)	0.1969(0.0165)	0.1969(0.0165)
$N = 100$	$T = 40$	0.2493(0.0120)	0.4871(0.0122)	0.2869(0.0349)	0.2851(0.0304)
	$T = 80$	0.1695(0.0082)	0.3606(0.0090)	0.1728(0.0108)	0.1728(0.0108)

fixed-design $X_{it} = t/T$. More specifically, data are generated from

$$Y_{it} = \beta_i\big(\frac{t}{T}\big) + \epsilon_{it},$$

where

$$\beta_i(u) = \begin{cases} \gamma_1(u) = G(u, \frac{1}{2}, \frac{1}{2}) & \text{if } i \in \mathcal{G}_1, \\ \gamma_2(u) = G(u, \frac{1}{4}, \frac{1}{4}) + G(u, \frac{3}{4}, \frac{1}{4}) & \text{if } i \in \mathcal{G}_2, \\ \gamma_3(u) = G(u, \frac{1}{8}, \frac{1}{8}) + G(u, \frac{3}{8}, \frac{1}{8}) + G(u, \frac{3}{4}, \frac{1}{4}) & \text{if } i \in \mathcal{G}_3, \\ \gamma_4(u) = G(u, \frac{1}{4}, \frac{1}{4}) + G(u, \frac{5}{8}, \frac{1}{8}) + G(u, \frac{7}{8}, \frac{1}{8}) & \text{if } i \in \mathcal{G}_4, \\ \gamma_5(u) = G(u, \frac{1}{12}, \frac{1}{12}) + G(u, \frac{1}{4}, \frac{1}{12}) + G(u, \frac{5}{12}, \frac{1}{12}) + G(u, \frac{3}{4}, \frac{1}{4}) & \text{if } i \in \mathcal{G}_5, \\ \gamma_6(u) = G(u, \frac{1}{4}, \frac{1}{4}) + G(u, \frac{7}{12}, \frac{1}{12}) + G(u, \frac{3}{4}, \frac{1}{12}) + G(u, \frac{11}{12}, \frac{1}{12}) & \text{if } i \in \mathcal{G}_6, \end{cases} \tag{4.5}$$

where

$$G(u, \mu, \nu) = I\big(\big|\frac{u-\mu}{\nu}\big| \le 1\big)\big[1 - \big(\frac{u-\mu}{\nu}\big)^2\big]^2,$$

the groups are defined as $\mathcal{G}_1 = \{1, 2, \ldots, N_1\}$, $\mathcal{G}_2 = \{N_1 + 1, \ldots, \sum_{k=1}^{2} N_k\}$, and $\mathcal{G}_3 = \{\sum_{k=1}^{2} N_k + 1, \ldots, \sum_{k=1}^{3} N_k\}$, $\mathcal{G}_4 = \{\sum_{k=1}^{3} N_k + 1, \ldots, \sum_{k=1}^{4} N_k\}$, $\mathcal{G}_5 = \{\sum_{k=1}^{4} N_k + 1, \ldots, \sum_{k=1}^{5} N_k\}$, and $\mathcal{G}_6 = \{\sum_{k=1}^{5} N_k + 1, \ldots, \sum_{k=1}^{6} N_k\}$, in which $N_k = N/6$, $k = 1, \ldots, 6$, ϵ_{it} are independently drawn from $N(0, \sigma^2)$ distribution

Table 4.4: Frequencies at which K_0 is estimated for Simulated Example 2

Error variance	GBIC					GAIC				
	4	5	6(true)	7	8	4	5	6(true)	7	8
$\sigma^2 = 0.49^2$ (NSR=2)	0	0	200	0	0	0	0	200	0	0
$\sigma^2 = 0.6^2$ (NSR=3)	1	0	199	0	0	1	0	199	0	0
$\sigma^2 = 0.7^2$ (NSR=4)	20	24	156	0	0	12	14	174	0	0

with $\sigma^2 = 0.49^2$, 0.60^2 and 0.7^2, which correspond to noise-to-signal ratios (NSR) of 2, 3, and 4. As in Vogt and Linton (2018), the sample size is set as $N = 240, T = 200$. The functions $\gamma_k(\cdot)$, $k = 3, 4, 5, 6$, have different smoothness in different regions of $[0, 1]$. Hence, a varying bandwidth (i.e., a bandwidth whose value varies with the point u at which $\beta_i(\cdot)$ is evaluated) may produce better estimation than a fixed-value bandwidth. However, for easier implementation, we still use a fixed bandwidth in the kernel estimation, which is selected via the cross-validation method detailed in Section 4.1. The subsequent clustering results (shown in Tables 4.4-4.6) are still satisfactory and comparable to those in Vogt and Linton (2018), which are obtained based on a distance measure that maximises over the domain of the coefficient functions and the range of values for the bandwidth h. However, our method is easier and more straightforward to implement.

As in Simulated Example 1, 200 repeated samples are drawn from the data generating process, and the same quantities (i.e., the frequencies at which the correct number of groups is chosen, the NMI and purity, and the RMSE of the functional coefficients estimation) are computed and presented in Tables 4.4-4.6. Unsurprisingly, as the error variance increases (or the NSR increases), the performance of both the GBIC and GAIC deteriorates, so does the accuracy of all the estimation approaches. However, even when the NSR is 4 ($\sigma^2 = 0.7^2$), the GAIC selects the correct number of groups in 87% of the replications and the GBIC in 78% of the replications. This number is around 82.5% in Vogt and Linton (2018) (although they have random-design X_{it} rather than fixed-design t/T as in our setting here). When the NSR is lower (i.e., 2 or 3), the GAIC and GBIC select the correct number of groups in almost all of the replications. The RMSE's of the post-clustering estimation of the functional coefficients for the GBIC and GAIC are close to that of the oracle estimation, and there is a reduction of around 45% in the RMSE by pooling data belonging to the same group, compared with the non-pooling pre-clustering estimation. We also note that while the sample size of $N = 240, T = 200$ in this example is larger than those in Simulated Example 1, the accuracy of the GBIC and GAIC for selecting K and the subsequence HAC results in this example is lower than that in Simulated Example 1. This is due to the fact that the NSRs in Simulated Example 1 are much smaller (between 0.18-0.30).

Table 4.5: The average(standard deviation) NMI's and purities for Simulated Example 2

Error variance	GBIC		GAIC	
	NMI	Purity	NMI	Purity
$\sigma^2 = 0.49^2$ (NSR=2)	0.9998(0.0013)	0.9999(0.0005)	0.9998(0.0013)	0.9999(0.0005)
$\sigma^2 = 0.6^2$ (NSR=3)	0.9933(0.0147)	0.9975(0.0043)	0.9933(0.0147)	0.9975(0.0043)
$\sigma^2 = 0.7^2$ (NSR=4)	0.9497(0.0530)	0.9879(0.0107)	0.9549(0.0468)	0.9850(0.0147)

Table 4.6: The average(standard deviation) RMSE's of $\beta_i(\cdot)$ estimates for Simulated Example 2

Error variance	Oracle	Pre-clustering	Post-clustering	
			GBIC	GAIC
$\sigma^2 = 0.49^2$ (NSR=2)	0.0527(0.0017)	0.1299(0.0015)	0.0527(0.0017)	0.0527(0.0017)
$\sigma^2 = 0.6^2$ (NSR=3)	0.0749(0.0019)	0.1517(0.0018)	0.0759(0.0048)	0.0759(0.0048)
$\sigma^2 = 0.7^2$ (NSR=4)	0.0818(0.0022)	0.1703(0.0020)	0.0939(0.0186)	0.0911(0.0154)

4.3 An empirical application

In this session we apply our kernel HAC method to a panel study of economic growth, in which we consider the following growth model

$$\mathrm{GY}_{it} = \alpha_i + \beta_{i,0}\big(\frac{t}{T}\big) + \beta_{i,1}\big(\frac{t}{T}\big)\mathrm{GK}_{it} + \beta_{i,2}\big(\frac{t}{T}\big)\mathrm{GPOP}_{it} + \epsilon_{it}, \quad i = 1, \ldots, N, t = 1, \ldots, T, \tag{4.6}$$

where GRY_{it} is the GDP annual growth rate of the i-th country in year t, GK_{it} is the annual growth rate of capital formation, and GPOP_{it} is the annual growth of population. All three variables are in percentages. Ideally, one would use the annual growth of labour input in place of GPOP_{it}, but since measures of labour input are scarce, we replace it with the annual population growth. The data are obtained from the World Bank's World Development Indicators (WDI) database and cover 61 countries over the period 1971–2016. A plot of the data for these variables is given in Figure 4.1.

We estimate the functional coefficients $\boldsymbol{\beta}_i(\,\cdot\,) = \big(\beta_{i,0}(\cdot), \beta_{i,1}(\cdot), \beta_{i,2}(\cdot)\big)'$ using nonparametric kernel smoothing with the Epanechnikov kernel and a bandwidth selected from the leave-one-out cross validation. Then, the kernel HAC method is used to classify the estimated $\boldsymbol{\beta}_i(\,\cdot\,)$ with the number of groups determined by the information criterion introduced in Section 2.2. Both GAIC and GBIC identify four groups with the estimated group-specific functional coefficients depicted in Figure 4.3. Figure 4.2 gives a dendrogram plot of the kernel HAC algorithm, in which the y-axis represents distance (measured as the "complete linkage") between groups of functional coefficient vectors, and the x-axis shows the indices of countries. The dendrogram consists of a series of U shapes, each

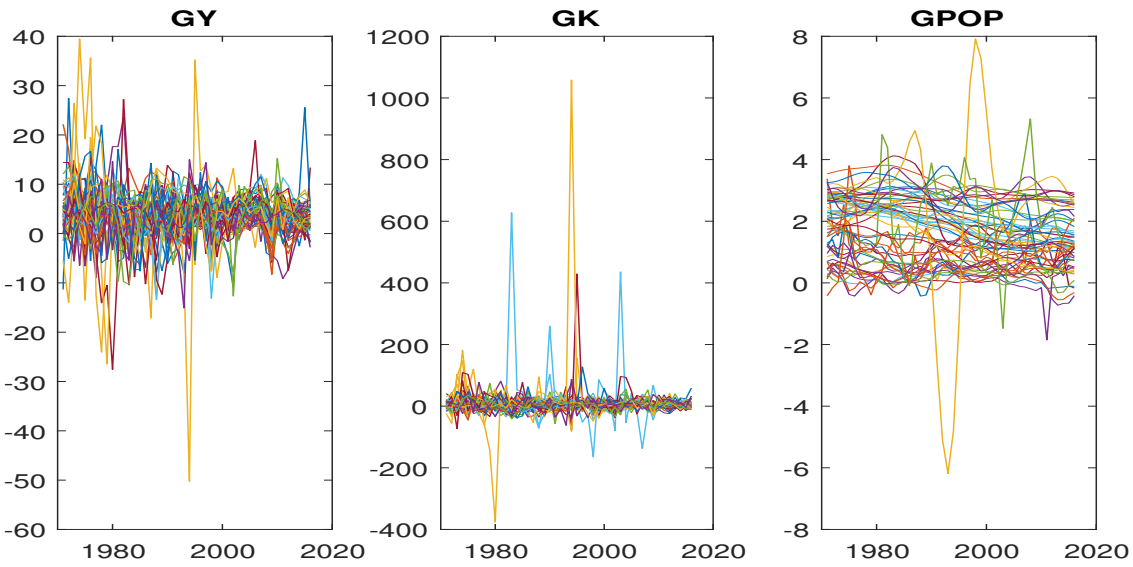

Figure 4.1: Plot of data. Left to right: GDP annual growth (in %), annual growth rate of capital formation (in %), annual growth rate of population (in %).

representing the joining of two groups in a hierarchical tree. The memberships of the four estimated groups are given in Table 4.7. A plot of the data by the 4 identified groups is given in Figure 4.4.

Most countries (48 out of 61 countries) are classified into the first group, while Groups 2 and 4 have 3 members each and Group 3 has 7 members. Figure 4.3 shows that the post-clustering estimates for the functional coefficients for Group 1 have smaller variations over the sample period than those for the other 3 groups do. This may indicate that the countries in Groups 2-4 experienced greater economic structural changes than the countries in Group 1. For all the groups, the growth of capital formation has an overall positive effect on the growth of GDP. However, the effect of population growth is mixed. For Group 1 this effect is mixed and for Group 3 it is mostly positive over the period considered. On the other hand, for Groups 2 and 4, it is mostly negative. Population growth for Group 2 countries has an increasing negative effect, while it has a decreasing negative effect for Group 4 countries.

5 Conclusions

In this paper we propose a kernel HAC method to estimate the latent group structure in a heterogeneous time-varying coefficient panel data model. This method applies the classic HAC method to the kernel estimates of functional coefficients from each cross section. It is easy to implement and provides a consistent estimate of the latent group structure when $T \to \infty$, irrespective of whether there is

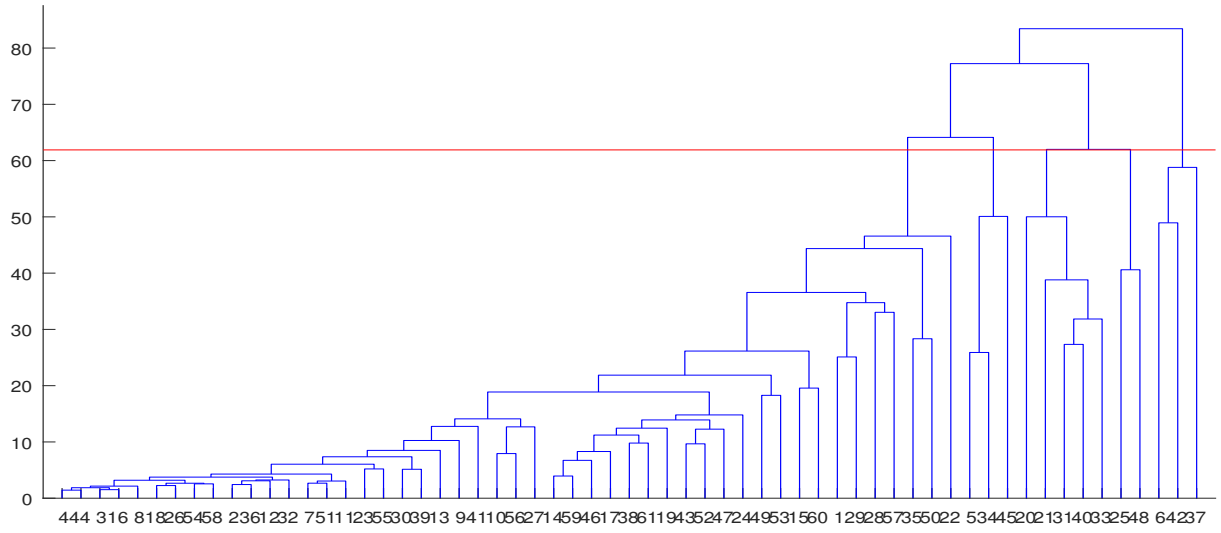

Figure 4.2: A dendrogram of the HAC algorithm

Table 4.7: Memberships of the four estimated groups

Groups	Countries
Group 1	1. Argentina, 2. Australia, 3. France, 4. Germany, 7. Italy, 8. Japan 9. Korea, Rep., 10. Mexico, 11. Netherlands, 12. Spain, 13. United Kingdom, 14. United States 15. Algeria, 16. Austria, 17. Bangladesh, 18. Belgium, 19. Benin, 22. Cameroon 23. Canada, 24. Colombia, 26. Denmark, 27. Dominican Republic, 28. Ecuador, 29. Egypt, Arab Rep. 30. Finland, 32. Greece, 35. Iran, Islamic Rep., 36. Ireland, 38. Lesotho, 39. Luxembourg 41. Malaysia, 43. Morocco, 44. New Zealand, 46. Norway, 47. Pakistan, 49. Peru 50. Philippines, 51. Portugal, 52. Rwanda, 53. Senegal, 54. Singapore, 55. South Africa 56. Sri Lanka, 57. Sudan, 58. Sweden, 59. Thailand, 60. Togo, 61. Uruguay
Group 2	5. India, 34. Honduras, 45. Nicaragua
Group 3	20. Bolivia, 21. Burkina Faso, 25. Congo, Rep, 31. Gabon, 33. Guatemala, 40. Madagascar, 48. Panama
Group 4	6. Indonesia, 37. Kenya, 42. Mauritania

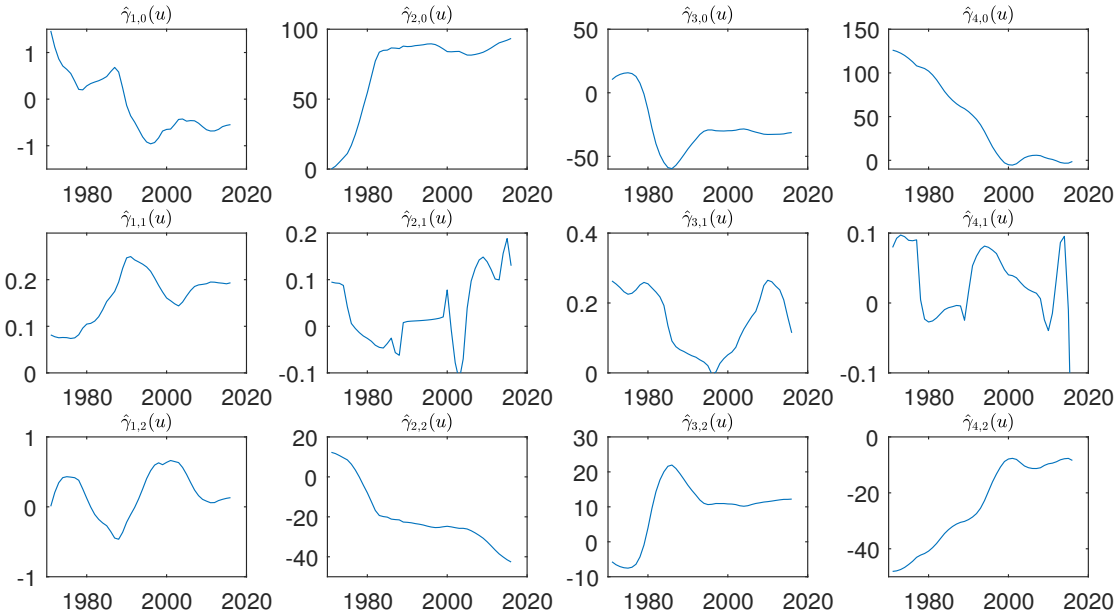

Figure 4.3: Post-clustering estimates of group-specific functional coefficients. Plots in each row represent a component of the estimated coefficient vector, one for each group.

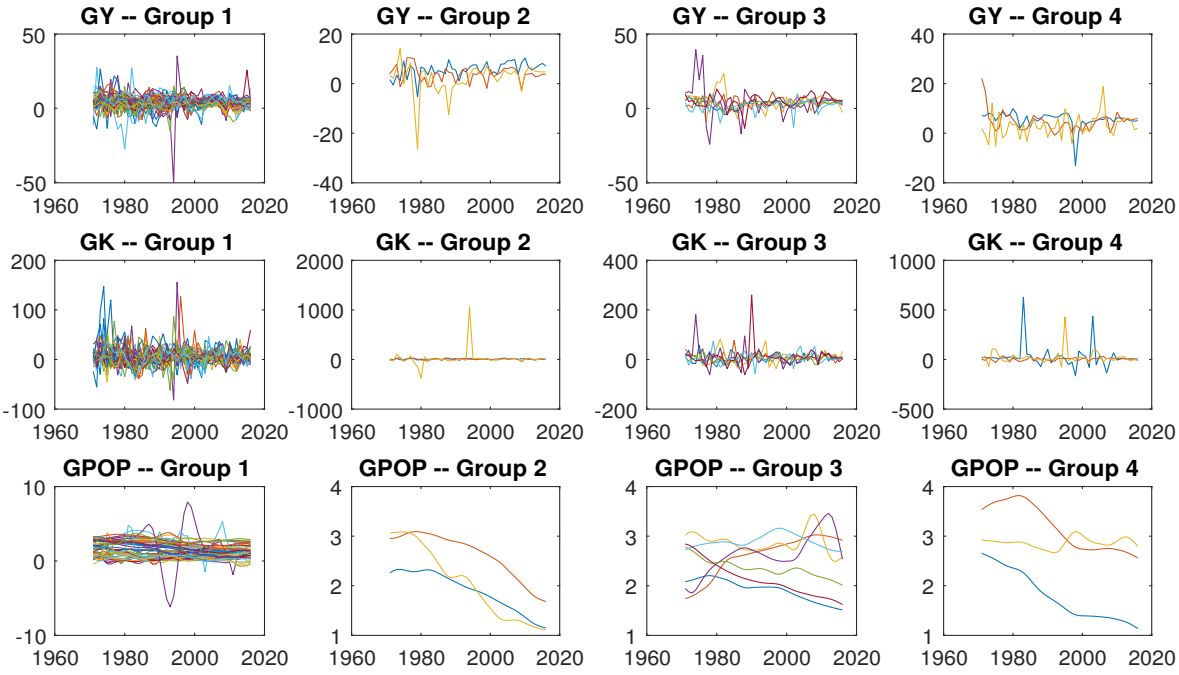

Figure 4.4: A plot of data by estimated groups.

cross-sectional dependence or not. We also introduce an information criterion to estimate the number of groups when it is unknown and propose two possible choices for the tuning parameter in the information criterion, which are then shown to work well in the simulation studies. The bandwidth used in the kernel estimation can be chosen via a data-driven method, such as the cross-validation method. In the simulation studies we adopt a data generating process from Su et al. (2018) and another from Vogt and Linton (2018) to see how our method performs in their settings. The results show that it performs comparably well to those of Su et al. (2018) and Vogt and Linton (2018). We also apply our method to a panel study of economic growth and identify 4 groups of countries which have different growth patterns.

A Technical conditions

We next list the technical assumptions which are needed to prove the main asymptotic results.

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

ASSUMPTION 2. *(i) For each i, the process $\{(X_{it,1}, \cdots, X_{it,p-1}, \epsilon_{it}) : 1 \leq t \leq T\}$ is stationary and α-mixing dependent with the mixing coefficient decaying to zero at a geometric rate.*

(ii) The explanatory variables $X_{it,j}$, $1 \leq j \leq p-1$, and disturbances ϵ_{it} satisfy the following moment conditions

$$\max_{1\leq i\leq N} \max_{1\leq j\leq p-1} \mathbb{E}\left(|X_{it,j}|^{2\delta}\right) < \infty, \quad \max_{1\leq i\leq N} \mathbb{E}\left(|\epsilon_{it}|^{2\delta}\right) < \infty, \tag{A.1}$$

where $\delta > 2(m+1)$ with m defiend in Assumption 4 below.

(iii) For each i, the $p \times p$ matrix $\mathbf{\Delta}_i = \mathbb{E}(\mathbf{X}_{it}\mathbf{X}_{it}')$ is positive definite. Furthermore, there exist two finite positive constants, $\underline{\lambda}$ and $\overline{\lambda}$, such that

$$0 < \underline{\lambda} \leq \min_{1\leq i\leq N} \lambda_{\min}\left(\mathbf{\Delta}_i\right) \leq \max_{1\leq i\leq N} \lambda_{\max}\left(\mathbf{\Delta}_i\right) \leq \overline{\lambda} < \infty, \tag{A.2}$$

where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimum and maximum eigenvalues of a square matrix, respectively.

ASSUMPTION 3. *(i) The group-specific coefficient functions $\boldsymbol{\gamma}_k(\cdot)$, $1 \leq k \leq K_0$, (and hence $\boldsymbol{\beta}_i(\cdot)$, $1 \leq i \leq N$), have continuous second-order derivatives on the interval $[0, 1]$.*

(ii) The weight function $W(\cdot)$ is non-negative and continuous on $[0, 1]$. In addition, there exists a small positive constant ω such that $W(u) = 0$ if $u \leq \omega$ or $u \geq 1 - \omega$.

ASSUMPTION 4. *(i) There exists a positive constant m such that $N = o(T^m)$.*

(ii) The bandwidth h satisfies $h \to 0$ and $(T^{1-2(m+1)/\delta}h)/\log^3 T \to \infty$, where δ was defined in Assumption 2(ii).

(iii) Letting

$$\zeta = \min_{1\le k\neq l\le K_0} \min_{i\in\mathcal{G}_k, j\in\mathcal{G}_l} \delta_{i,j},$$

we have $h^2 + [\log T/(Th)]^{1/2} = o(\zeta)$.

ASSUMPTION 5 *(i) There exist two positive constants τ_1, with $0 < \tau_1 < 1$, and τ_2 such that*

$$\min_{1\le k\le K_0} |\mathcal{G}_k| \ge \tau_1 \cdot N, \quad \min_{1\le k_1\neq k_2\le K_0} \int_{\omega}^{1-\omega} \left\|\boldsymbol{\gamma}_{k_1}(u) - \boldsymbol{\gamma}_{k_2}(u)\right\|_2^2 W(u)du > \tau_2. \tag{A.3}$$

(ii) The tuning parameter ρ satisfies $\rho \to 0$ and $\frac{\log T}{T} + h^4 + \frac{1}{NTh} = o(\rho)$.

ASSUMPTION 6. *For any index set $\mathcal{G} \subset \mathcal{G}_k$, $k = 1, \cdots, K_0$,*

$$\sum_{i\in\mathcal{G}} \mathbb{E}\left[\sum_{t=1}^{T} \|\mathbf{X}_{it}\|_2^2 \cdot \left\|\sum_{j\in\mathcal{G}}\sum_{s=1}^{T} \epsilon_{js}\mathbf{X}_{js}K_{st}\right\|_2^2\right] = O\left(|\mathcal{G}|^2T^2h\right) \tag{A.4}$$

and

$$\mathbb{E}\left(\left[\sum_{i\in\mathcal{G}}\sum_{t=1}^{T}\sum_{j\in\mathcal{G}}\sum_{s=1}^{T} \epsilon_{it}\epsilon_{js}K_{st}\mathbf{X}_{it}'\left(\frac{1}{|\mathcal{G}|}\sum_{l\in\mathcal{G}}\boldsymbol{\Delta}_l\right)^{-1}\mathbf{X}_{js}\right]^2\right) = O\left(|\mathcal{G}|^2T^2\right) \tag{A.5}$$

where $K_{st} = K\left(\frac{s-t}{Th}\right)$ and $\boldsymbol{\Delta}_i$ was defined in Assumption 2(iii).

REMARK A.1. The conditions on the kernel function $K(\cdot)$ in Assumption 1 are mild and satisfied by some commonly-used kernel functions such as the Epanechnikov kernel and uniform kernel. Assumption 2 allows that the panel time series observations are temporally correlated and the α-mixing dependence is one of the weakest dependence conditions. The moment conditions in (A.1) and (A.2) are crucial to derive uniform convergence (uniform over i and u) of some kernel-based quantities. The smoothness conditions on the coefficient functions and weight function in Assumption 3 are not uncommon. In particular, Assumption 3(ii) indicates that the kernel estimates are truncated at those scaled time points that are close to the boundaries (0 and 1). Assumption 4 imposes some mild restriction on the bandwidth, the relationship between the cross-sectional size and time series length, and the smallest L_q-distance between coefficient functions for different groups. A combination of Assumptions 2(ii) and 4(i) indicates that there is a trade-off between the moment conditions and the divergence rate of N. If the cross-sectional size diverges at a faster rate (m becomes larger), stronger moment conditions (i.e., larger δ) would be required for the relevant asymptotic theory. In fact,

our theory still holds when N diverges at an exponential rate of T. In the latter case, exponential moment conditions would be needed for $\mathbf{X}_{it}$ and ϵ_{it}. In addition, when δ is very large, the restriction on the bandwidth in Assumption 4(ii) would become weaker. Assumption 4(iii) indicates that ζ can converge to zero at an appropriate rate. Assumptions 5 and 6 are mainly used to prove consistency of $\widehat{K}$ from the information criterion proposed in Section 2.2. Assumption 5(i) is crucial in order to show that $\mathbb{IC}(K) > \mathbb{IC}(K_0)$ when $K < K_0$ (i.e., the model is under-identified). The high-order moment conditions in Assumption 6 indicate that the panel observations can be serially correlated and weakly cross-sectionally dependent, and both (A.4) and (A.5) are easy to verify when $X_{it,j}$ and ϵ_{it} are independent over both i and t.

B Proofs of the asymptotic results

In this appendix, we give the detailed proofs of the main theoretical results in Section 3. We start with a proposition which shows the convergence rates of the individual functional coefficient estimators $\widehat{\boldsymbol{\beta}}_i(u)$ (defined in Section 2.1) uniformly over i and u, without placing any restrictions on the panel cross-sectional dependence.

PROPOSITION B.1. *Let Assumptions 1, 2, 3(i) and 4(i)(ii) in Appendix A hold. Then, as $T \to \infty$, we have*

$$\max_{1\le i\le N} \sup_{\omega\le u\le 1-\omega} \left\|\widehat{\boldsymbol{\beta}}_i(u) - \boldsymbol{\beta}_i(u)\right\|_q = O_P\left(h^2 + \eta_1(T,h)\right), \tag{B.1}$$

where $q \ge 1$, $\eta_1(T,h) = [\log T/(Th)]^{1/2}$ and ω is a small positive constant defined in Assumption 3(ii).

PROOF OF PROPOSITION B.1. In this proof and that of Theorem 1, all the limiting results are established under $T \to \infty$.

By the definition of $\widehat{\boldsymbol{\beta}}_i(u)$, we only need to show

$$\max_{1\le i\le N} \sup_{h\le u\le 1-h} \left\|\widehat{\boldsymbol{\beta}}^*_i(u) - \boldsymbol{\beta}^*_i(u)\right\|_q = O_P\left(h^2 + \eta_1(T,h)\right), \tag{B.2}$$

and

$$\max_{1\le i\le N} \sup_{\omega\le u\le 1-\omega} \left|\widehat{\beta}_{i,0}(u) - \beta_{i,0}(u)\right| = O_P\left(h^2 + \eta_1(T,h)\right), \tag{B.3}$$

where $\widehat{\boldsymbol{\beta}}^*_i(u)$ is defined in (2.1) and $\widehat{\beta}_{i,0}(u)$ is defined similarly to $\widehat{\boldsymbol{\beta}}^*_i(u)$ but with $\mathbf{X}_{it}$ and Y_{it} replaced by 1 and $\widehat{Z}^c_{it}$ (defined in Section 2.1), respectively.

PROOF OF (B.2). Letting $K_t(u) = K\left(\frac{t-uT}{Th}\right)$, we note that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_i^*(u) - \boldsymbol{\beta}_i^*(u) &= \left[\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{it}' K_t(u)\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{X}_{it} Y_{it} K_t(u)\right] - \boldsymbol{\beta}_i^*(u) \\
&= \left[\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{it}' K_t(u)\right]^{-1} \left\{\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{it}' \left[\boldsymbol{\beta}_i^*(t/T) - \boldsymbol{\beta}_i^*(u)\right] K_t(u)\right\} + \\
&\quad \left[\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{it}' K_t(u)\right]^{-1} \left[\sum_{t=1}^{T} \mathbf{X}_{it}\epsilon_{it} K_t(u)\right] \\
&=: \boldsymbol{\Lambda}_{i,1}(u) + \boldsymbol{\Lambda}_{i,2}(u). \qquad \text{(B.4)}
\end{aligned}$$

To prove (B.2), it is sufficient to show

$$\max_{1\le i\le N} \sup_{h\le u\le 1-h} \left\|\boldsymbol{\Lambda}_{i,1}(u)\right\|_q = O_P(h^2) \qquad \text{(B.5)}$$

and

$$\max_{1\le i\le N} \sup_{h\le u\le 1-h} \left\|\boldsymbol{\Lambda}_{i,2}(u)\right\|_q = O_P\left(\eta_1(T,h)\right). \qquad \text{(B.6)}$$

We first give the detailed proof of (B.6) and then sketch the proof of (B.5). Observe that

$$\begin{aligned}
&\frac{1}{Th}\sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}_{it}' K_t(u) - \boldsymbol{\Delta}_i \\
=\ &\frac{1}{Th}\sum_{t=1}^{T} \left[\mathbf{X}_{it}\mathbf{X}_{it}' - \boldsymbol{\Delta}_i\right] K_t(u) + \boldsymbol{\Delta}_i \left[\frac{1}{Th}\sum_{t=1}^{T} K_t(u) - 1\right] \\
=\ &\frac{1}{Th}\sum_{t=1}^{T} \left[\mathbf{X}_{it}\mathbf{X}_{it}' - \boldsymbol{\Delta}_i\right] K_t(u) + O\left(1/(Th)\right), \qquad \text{(B.7)}
\end{aligned}$$

as it is easy to show, by Assumption 1, that

$$\frac{1}{Th}\sum_{t=1}^{T} K_t(u) = \int_{-1}^{-1} K(w)dw + O\left(1/(Th)\right) = 1 + O\left(1/(Th)\right) \qquad \text{(B.8)}$$

uniformly over $h \le u \le 1-h$.

Letting $Q_{it}(X) = \mathbf{X}_{it}\mathbf{X}_{it}' - \boldsymbol{\Delta}_i$ and $q_{it}^{(j,k)}(X)$ be the (j,k)-entry of $Q_{it}(X)$, we next prove that

$$\max_{1\le j,k\le p} \max_{1\le i\le N} \sup_{h\le u\le 1-h} \left|\frac{1}{Th}\sum_{t=1}^{T} q_{it}^{(j,k)}(X) K_t(u)\right| = O_P\left(\eta_1(T,h)\right). \qquad \text{(B.9)}$$

As the number p is fixed, it is sufficient to prove

$$\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left|\frac{1}{Th}\sum_{t=1}^{T}q_{it}^{(j,k)}(X)K_t(u)\right| = O_P\left(\eta_1(T,h)\right) \tag{B.10}$$

for each (j,k). To prove (B.10), we use the truncation technique and define

$$\overline{q}_{it}^{(j,k)}(X) = q_{it}^{(j,k)}(X)I\left(|q_{it}^{(j,k)}(X)|\le T^{(m+1)/\delta}\right),\quad \widetilde{q}_{it}^{(j,k)}(X) = q_{it}^{(j,k)}(X) - \overline{q}_{it}^{(j,k)}(X),$$

where $I(\cdot)$ denotes the indicator function, δ and m were defined in Assumptions 2 and 4, respectively.

By Assumptions 2(ii) and 4(i), and the Bonferroni and Markov inequalities, we may show that for any $\xi>0$,

$$\begin{aligned}
& \mathbb{P}\left(\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left|\frac{1}{Th}\sum_{t=1}^{T}\widetilde{q}_{it}^{(j,k)}(X)K_t(u)\right| > \xi\eta_1(T,h)\right)\\
\le\ & \mathbb{P}\left(\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left|\frac{1}{Th}\sum_{t=1}^{T}\widetilde{q}_{it}^{(j,k)}(X)K_t(u)\right| > 0\right)\\
\le\ & \mathbb{P}\left(\max_{1\le i\le N}\max_{1\le t\le T}\left|q_{it}^{(j,k)}(X)\right| > T^{(m+1)/\delta}\right)\\
\le\ & \sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{P}\left(\left|q_{it}^{(j,k)}(X)\right| > T^{(m+1)/\delta}\right)\\
\le\ & N\cdot T\max_{1\le i\le N}\max_{1\le t\le T}\mathbb{E}\left[\left|q_{it}^{(j,k)}(X)\right|^{\delta}\right]/T^{m+1}\\
=\ & O\left(N/T^m\right) = o(1),
\end{aligned}$$

which leads to

$$\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left|\frac{1}{Th}\sum_{t=1}^{T}\widetilde{q}_{it}^{(j,k)}(X)K_t(u)\right| = o_P\left(\eta_1(T,h)\right). \tag{B.11}$$

On the other hand, we consider covering the closed interval $[h,1-h]$ by some disjoint sub-intervals $\mathcal{I}_l$, $1\le l\le L$, with centres u_l and length $h\eta_1(T,h)/T^{(m+1)/\delta}$. It is easy to calculate that the number

of such sub-intervals, L, is bounded by $T^{(m+1)/\delta}/\left[h\eta_1(T,h)\right]$. Note that

$$\begin{aligned}
&\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left|\frac{1}{Th}\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)K_t(u)\right| \\
\le\ &\max_{1\le i\le N}\max_{1\le l\le L}\left|\frac{1}{Th}\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)K_t(u_l)\right| + \max_{1\le i\le N}\max_{1\le l\le L}\sup_{u\in\mathcal{I}_l}\left|\frac{1}{Th}\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)\left[K_t(u)-K_t(u_l)\right]\right| \\
=\ &\max_{1\le i\le N}\max_{1\le l\le L}\left|\frac{1}{Th}\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)K_t(u_l)\right| + O_P\left(\eta_1(T,h)\right) \qquad \text{(B.12)}
\end{aligned}$$

by using Assumption 1 and the definition of $\overline{q}_{it}^{(j,k)}(X)$.

Finally, by the Bonferroni inequality again and the exponential-type inequality for α-mixing sequences (e.g., Lemma 1.3(2) in Bosq, 1998), we may show that

$$\begin{aligned}
&\mathbb{P}\left(\max_{1\le i\le N}\max_{1\le l\le L}\left|\frac{1}{Th}\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)K_t(u_l)\right| > \xi\eta_1(T,h)\right) \\
\le\ &\sum_{i=1}^{N}\sum_{l=1}^{L}\mathbb{P}\left(\left|\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)K_t(u_l)\right| > \xi\cdot(Th)\cdot\eta_1(T,h)\right) \\
=\ &O\left(NL\exp\{-c_1\xi^2\log T\}\right) + O\left(NL\cdot\frac{Th\gamma_0^{c_2\log T}}{\eta_1^{3/2}(T,h)T^{(m+1)/(2\delta)}}\right), \qquad \text{(B.13)}
\end{aligned}$$

where c_1 is a fixed positive constant, $0<\gamma_0<1$, ξ and c_2 are chosen to be sufficiently large so that the orders on the right hand side of the equality in (B.13) becomes $o(1)$ (noting that both N and L diverge at certain polynomial rate of T). Combining (B.12) and (B.13), we readily have

$$\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left|\frac{1}{Th}\sum_{t=1}^{T}\overline{q}_{it}^{(j,k)}(X)K_t(u)\right| = O_P\left(\eta_1(T,h)\right). \qquad \text{(B.14)}$$

With (B.11) and (B.14), we prove (B.9), which together with (B.7), leads to

$$\frac{1}{Th}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{it}'K_t(u) = \mathbf{\Delta}_i + O_P\left(\eta_1(T,h)\right) \qquad \text{(B.15)}$$

uniformly over $1\le i\le N$ and $h\le u\le 1-h$. Following similar arguments in the proof of (B.15) and noting that $\mathbb{E}(\epsilon_{it}|\mathbf{X}_{it})=0$ a.s., we can prove that

$$\max_{1\le i\le N}\sup_{h\le u\le 1-h}\left\|\frac{1}{Th}\sum_{t=1}^{T}\mathbf{X}_{it}\epsilon_{it}K_t(u)\right\|_q = O_P\left(\eta_1(T,h)\right). \qquad \text{(B.16)}$$

Using (B.15), (B.16) and Assumption 2(iii), we can complete the proof of (B.6).

On the other hand, by Assumption 3(i) and using the Taylor expansion for $\boldsymbol{\beta}_i^*(\cdot)$, we may show that (B.5) holds. With (B.5) and (B.6), we complete the proof of (B.2). □

Proof of (B.3). Recall that

$$\widehat{\beta}_{i,0}(u) = \sum_{t=1}^{T} \widehat{Z}_{it}^c K_t(u) / \sum_{t=1}^{T} K_t(u).$$

Let

$$\widetilde{\beta}_{i,0}(u) = \sum_{t=1}^{T} Z_{it}^c K_t(u) / \sum_{t=1}^{T} K_t(u)$$

be an infeasible kernel estimate of $\beta_{i,0}(u)$ with Z_{it}^c defined as

$$Z_{it}^c = Z_{it} - \frac{1}{T}\sum_{s=1}^{t} Z_{is}, \quad Z_{it} = Y_{it} - \sum_{k=1}^{p-1} X_{it,k}\beta_{i,k}\big(\frac{t}{T}\big).$$

Then, in order to prove (B.3), we need only to show

$$\max_{1\le i\le N} \sup_{\omega\le u\le 1-\omega} \left|\widetilde{\beta}_{i,0}(u) - \beta_{i,0}(u)\right| = O_P\left(h^2 + \eta_1(T,h)\right) \tag{B.17}$$

and

$$\max_{1\le i\le N} \sup_{\omega\le u\le 1-\omega} \left|\widehat{\beta}_{i,0}(u) - \widetilde{\beta}_{i,0}(u)\right| = O_P\left(h^2 + \eta_1(T,h)\right). \tag{B.18}$$

Note that

$$Z_{it}^c = Z_{it} - \frac{1}{T}\sum_{s=1}^{t} Z_{is} = \beta_{i,0}\big(\frac{t}{T}\big) + \epsilon_{it} - \frac{1}{T}\sum_{s=1}^{T} \epsilon_{is} =: \beta_{i,0}\big(\frac{t}{T}\big) + \epsilon_{it}^c,$$

and

$$\begin{aligned}\widetilde{\beta}_{i,0}(u) - \beta_{i,0}(u) &= \sum_{t=1}^{T} \epsilon_{it}^c K_t(u) / \sum_{t=1}^{T} K_t(u) + \left[\sum_{t=1}^{T} \beta_{i,0}\big(\frac{t}{T}\big) K_t(u) / \sum_{t=1}^{T} K_t(u) - \beta_{i,0}(u)\right] \\ &=: \Lambda_{i,3}(u) + \Lambda_{i,4}(u).\end{aligned} \tag{B.19}$$

For $\Lambda_{i,3}(u)$, we may decompose it as

$$\Lambda_{i,3}(u) = \sum_{t=1}^{T} \epsilon_{it} K_t(u) / \sum_{t=1}^{T} K_t(u) - \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it} =: \Lambda_{i,5}(u) + \Lambda_{i,6}. \tag{B.20}$$

Following the proof of (B.15), we may prove that

$$\max_{1\le i\le N}\sup_{\omega\le u\le 1-\omega}|\Lambda_{i,5}(u)| = O_P\left(\eta_1(T,h)\right) \tag{B.21}$$

and

$$\max_{1\le i\le N}|\Lambda_{i,6}| = \max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\right| = O_P\left(\eta_2(T)\right) \tag{B.22}$$

with $\eta_2(T) = (\log T/T)^{1/2}$. In view of (B.20)–(B.22), we readily have

$$\max_{1\le i\le N}\sup_{\omega\le u\le 1-\omega}|\Lambda_{i,3}(u)| = O_P\left(\eta_1(T,h)\right). \tag{B.23}$$

On the other hand, by Assumption 3(i) and using the Taylor expansion of $\beta_{i,0}(\cdot)$, we may prove that

$$\max_{1\le i\le N}\sup_{\omega\le u\le 1-\omega}|\Lambda_{i,4}(u)| = O_P(h^2), \tag{B.24}$$

which, together with (B.19) and (B.23), leads to (B.17).

We next consider the proof of (B.18). Observe that

$$\begin{aligned}\widehat{\beta}_{i,0}(u) - \widetilde{\beta}_{i,0}(u) &= \sum_{t=1}^{T}\left(\widehat{Z}_{it} - Z_{it}\right)K_t(u)/\sum_{t=1}^{T}K_t(u) - \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{Z}_{it} - Z_{it}\right)\\ &=: \Lambda_{i,7}(u) + \Lambda_{i,8}.\end{aligned} \tag{B.25}$$

By (B.2) with $q = 2$ and noting that the kernel function $K(\cdot)$ has a compact support $[-1,1]$ (see Assumption 1), we may show that

$$\begin{aligned}&\max_{1\le i\le N}\sup_{\omega\le u\le 1-\omega}\left|\frac{1}{Th}\sum_{t=1}^{T}\left(\widehat{Z}_{it} - Z_{it}\right)K_t(u)\right|\\ =\ &\max_{1\le i\le N}\sup_{\omega\le u\le 1-\omega}\left|\frac{1}{Th}\sum_{t=\lfloor(u-h)T\rfloor}^{\lfloor(u+h)T\rfloor+1}\left(\widehat{Z}_{it} - Z_{it}\right)K_t(u)\right|\\ =\ &c_3\cdot\max_{1\le i\le N}\sup_{\omega\le u\le 1-\omega}\left(\sum_{k=1}^{p-1}\left|\widehat{\beta}_{i,k}(u) - \beta_{i,k}(u)\right|^2\right)^{1/2}\\ =\ &O_P\left(h^2 + \eta_1(T,h)\right),\end{aligned} \tag{B.26}$$

where $\lfloor\cdot\rfloor$ denotes the floor function and c_3 is a positive constant.

On the other hand, we note that

$$\Lambda_{i,8} = \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{Z}_{it} - Z_{it}\right) = \frac{1}{T}\left(\sum_{t=1}^{\lfloor Th\rfloor} + \sum_{t=\lfloor Th\rfloor+1}^{T-\lfloor Th\rfloor} + \sum_{t=T-\lfloor Th\rfloor+1}^{T}\right)\left(\widehat{Z}_{it} - Z_{it}\right). \tag{B.27}$$

Following the proof of (B.2), we may show that

$$\max_{1\le i\le N}\sup_{0\le u\le h}\left(\sum_{k=1}^{p-1}\left|\widehat{\beta}_{i,k}(u) - \beta_{i,k}(u)\right|^2\right)^{1/2} = O_P\left(h + \eta_1(T,h)\right). \tag{B.28}$$

The uniform convergence rate in (B.28) is slower than that in (B.2) due to the kernel estimation boundary effect. Similarly, the uniform consistency result still holds if $\sup_{0\le u\le h}$ in (B.28) is replaced by $\sup_{1-h\le u\le 1}$. Consequently, we can prove that

$$\max_{1\le i\le N}\left|\frac{1}{T}\left(\sum_{t=1}^{\lfloor Th\rfloor} + \sum_{t=T-\lfloor Th\rfloor+1}^{T}\right)\left(\widehat{Z}_{it} - Z_{it}\right)\right| = O_P\left(h(h + \eta_1(T,h))\right). \tag{B.29}$$

Similarly to the proof of (B.26), we have

$$\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=\lfloor Th\rfloor+1}^{T-\lfloor Th\rfloor}\left(\widehat{Z}_{it} - Z_{it}\right)\right| = O_P\left(h^2 + \eta_1(T,h)\right). \tag{B.30}$$

With (B.27), (B.29) and (B.30), we have

$$\max_{1\le i\le N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{Z}_{it} - Z_{it}\right)\right| = O_P\left(h^2 + \eta_1(T,h)\right), \tag{B.31}$$

which, together with (B.8), (B.25) and (B.26), leads to (B.18). The proof of (B.3) has been completed. □

We next make use of Proposition B.1 to prove Theorem 1.

Proof of Theorem 1. Let

$$\widetilde{\delta}_{ij} = \frac{1}{T}\sum_{t=1}^{T}\left\|\boldsymbol{\beta}_i(t/T) - \boldsymbol{\beta}_j(t/T)\right\|_q W(t/T). \tag{B.32}$$

To prove (3.1) in Theorem 1, we only need to show that

$$\mathbb{P}\left(\max_{1\le k\le K_0}\max_{i,j\in\mathcal{G}_k}\widehat{\delta}_{i,j} < \min_{1\le k\ne l\le K_0}\min_{i\in\mathcal{G}_k,j\in\mathcal{G}_l}\widehat{\delta}_{i,j}\right)\to 1 \tag{B.33}$$

as T tends to infinity. Note that for the distance between true functional coefficients, we have $\delta_{ij}\equiv 0$ if $i,j\in\mathcal{G}_k$, and

$$\min_{1\le k\ne l\le K_0}\min_{i\in\mathcal{G}_k,j\in\mathcal{G}_l}\delta_{i,j}=\zeta>0,$$

where ζ is defined in Assumption 4(iii). Hence, to prove (B.33), it is sufficient to prove that

$$\max_{1\le i,j\le N}\left|\widehat{\delta}_{ij}-\delta_{ij}\right| = o_P(\zeta). \tag{B.34}$$

Note that

$$\left|\widehat{\delta}_{ij}-\delta_{ij}\right|\le\left|\widehat{\delta}_{ij}-\widetilde{\delta}_{ij}\right|+\left|\widetilde{\delta}_{ij}-\delta_{ij}\right|.$$

By Assumption 3 and the definition of Riemann integral, we readily have that

$$\max_{1\le i,j\le N}\left|\widetilde{\delta}_{ij}-\delta_{ij}\right| = O_P(1/T)=o_P(\zeta). \tag{B.35}$$

On the other hand, by the Minkowski inequality and Proposition B.1, we may prove that uniformly over $1\le i,j\le N$,

$$\begin{aligned}
\left|\widehat{\delta}_{ij}-\widetilde{\delta}_{ij}\right| &\le \frac{1}{T}\sum_{t=1}^{T}\left|\left\|\widehat{\boldsymbol{\beta}}_i(t/T)-\widehat{\boldsymbol{\beta}}_j(t/T)\right\|_q-\left\|\boldsymbol{\beta}_i(t/T)-\boldsymbol{\beta}_j(t/T)\right\|_q\right|W(t/T)\\
&= \frac{1}{T}\sum_{t=\lfloor T\omega\rfloor}^{T-\lfloor T\omega\rfloor+1}\left|\left\|\widehat{\boldsymbol{\beta}}_i(t/T)-\widehat{\boldsymbol{\beta}}_j(t/T)\right\|_q-\left\|\boldsymbol{\beta}_i(t/T)-\boldsymbol{\beta}_j(t/T)\right\|_q\right|W(t/T)\\
&\le \frac{1}{T}\sum_{t=\lfloor T\omega\rfloor}^{T-\lfloor T\omega\rfloor+1}\left|\left[\left\|\widehat{\boldsymbol{\beta}}_i(t/T)-\boldsymbol{\beta}_i(t/T)\right\|_q+\left\|\widehat{\boldsymbol{\beta}}_j(t/T)-\boldsymbol{\beta}_j(t/T)\right\|_q\right]W(t/T)\right|\\
&= O_P\left(h^2+\eta_1(T,h)\right).
\end{aligned} \tag{B.36}$$

Then, by (B.35), (B.36) and the triangle inequality, we can prove (B.34) by noting that $h^2+\eta_1(T,h)=o(\zeta)$ in Assumption 4(iii). The proof of Theorem 1 has been completed. □

We next provide the detailed proof of Theorem 2.

Proof of Theorem 2. From the definition of $\widehat{K}$ in (2.10), we only need to show that

$$\mathbb{P}\left(\mathbb{IC}(K_0)=\min_{1\le K\le\widetilde{K}}\mathbb{IC}(K)\right)\to 1. \tag{B.37}$$

Consider the following two cases: (i) $1 \leq K \leq K_0 - 1$ and (ii) $K_0 + 1 \leq K \leq \widetilde{K}$, which correspond, respectively, to under-identification and over-identification of the latent groups. Let $\mathcal{M}(\mathcal{G})$ denote the event that $\{\widehat{\mathcal{G}}_1, \cdots, \widehat{\mathcal{G}}_{K_0}\} = \{\mathcal{G}_1, \cdots, \mathcal{G}_{K_0}\}$. For case (i), by Theorem 1 and Lemma B.1 below, we have

$$\begin{aligned} & \mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K), 1 \leq K \leq K_0 - 1\right) \\ = \; & \mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K), 1 \leq K \leq K_0 - 1, \mathcal{M}(\mathcal{G})\right) + o(1) \\ = \; & 1 + o(1). \end{aligned} \tag{B.38}$$

On the other hand, for case (ii), by Theorem 1 and Lemma B.2 below, we have

$$\begin{aligned} & \mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K), K_0 + 1 \leq K \leq \widetilde{K}\right) \\ = \; & \mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K), K_0 + 1 \leq K \leq \widetilde{K}, \mathcal{M}(\mathcal{G})\right) + o(1) \\ = \; & 1 + o(1). \end{aligned} \tag{B.39}$$

Combining (B.38) and (B.39), we complete the proof of Theorem 2. □

LEMMA B.1. *Suppose that the assumptions in Theorem 2 are satisfied. Then we have*

$$\mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K), 1 \leq K \leq K_0 - 1\right) \to 1 \tag{B.40}$$

conditional on the event $\mathcal{M}(\mathcal{G})$.

PROOF OF LEMMA B.1. Without loss of generality, we only consider the case of $K = K_0 - 1$ as the other cases can be dealt with in the same manner. Conditional on the event $\mathcal{M}(\mathcal{G})$, two of the clusters among $\mathcal{G}_1, \cdots, \mathcal{G}_{K_0}$ are falsely merged when the HAC algorithm stops at $K = K_0 - 1$. Without loss of generality, we assume that $\mathcal{G}_{K_0-1}$ and $\mathcal{G}_{K_0}$ are falsely merged, and let $\boldsymbol{\gamma}_{k|K_0-1}(\cdot) = \boldsymbol{\gamma}_k(\cdot)$ for $k = 1, \cdots, K_0 - 2$, and $\boldsymbol{\gamma}_{K_0-1|K_0-1}(\cdot)$ be the vector of "pseudo" functional coefficients associated with $\mathcal{G}_{K_0-1|K_0-1} := \mathcal{G}_{K_0-1} \bigcup \mathcal{G}_{K_0}$

For $i \in \mathcal{G}_k$, $k = 1, \cdots, K_0 - 2$, note that

$$\begin{aligned} Y_{it}^c - \mathbf{X}_{it}' \widehat{\boldsymbol{\gamma}}_{k|K_0-1}(t/T) &= Y_{it} - \frac{1}{T}\sum_{s=1}^{T} \widehat{Z}_{is} - \mathbf{X}_{it}' \widehat{\boldsymbol{\gamma}}_{k|K_0-1}(t/T) \\ &= Y_{it} - \frac{1}{T}\sum_{s=1}^{T} Z_{is} + \frac{1}{T}\sum_{s=1}^{T}\left(Z_{is} - \widehat{Z}_{is}\right) - \mathbf{X}_{it}' \widehat{\boldsymbol{\gamma}}_{k|K_0-1}(t/T) \\ &= \epsilon_{it} - \frac{1}{T}\sum_{s=1}^{T} \epsilon_{is} + \frac{1}{T}\sum_{s=1}^{T}\left(Z_{is} - \widehat{Z}_{is}\right) - \mathbf{X}_{it}' \left[\widehat{\boldsymbol{\gamma}}_{k|K_0-1}(t/T) - \boldsymbol{\gamma}_k(t/T)\right]. \end{aligned}$$

Following the proof of Proposition B.1, we may show that

$$\max_{1\le k\le K_0-2}\max_{\omega\le u\le 1-\omega}\left\|\widehat{\boldsymbol{\gamma}}_{k|K_0-1}(u)-\boldsymbol{\gamma}_k(u)\right\|_2=O_P\left(h^2+\eta_1(T,h)\right). \tag{B.41}$$

This, together with (B.22) and (B.31), indicates that

$$\frac{1}{NT}\sum_{k=1}^{K_0-2}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\left[Y_{it}^c-\mathbf{X}_{it}'\widehat{\boldsymbol{\gamma}}_{k|K}(t/T)\right]^2W(t/T)=\frac{1}{NT}\sum_{k=1}^{K_0-2}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\epsilon_{it}^2W(t/T)+o_P(1). \tag{B.42}$$

For $i\in\mathcal{G}_{K_0-1}$, we note that

$$Y_{it}^c-\mathbf{X}_{it}'\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(t/T)=\epsilon_{it}-\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}+\frac{1}{T}\sum_{s=1}^{T}\left(Z_{is}-\widehat{Z}_{is}\right)-\mathbf{X}_{it}'\left[\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(t/T)-\boldsymbol{\gamma}_{K_0-1}(t/T)\right]$$

and conditional on $\mathcal{M}(\mathcal{G})$,

$$\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(u)=\left[\sum_{i\in\mathcal{G}_{K_0-1|K_0-1}}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{it}'K\left(\frac{t-uT}{Th}\right)\right]^{-1}\left[\sum_{i\in\mathcal{G}_{K_0-1|K_0-1}}\sum_{t=1}^{T}\mathbf{X}_{it}Y_{it}^cK\left(\frac{t-uT}{Th}\right)\right],$$

where $\mathcal{G}_{K_0-1|K_0-1}=\mathcal{G}_{K_0-1}\bigcup\mathcal{G}_{K_0}$ and

$$Y_{it}^c=\begin{cases}\epsilon_{it}-\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}+\frac{1}{T}\sum_{s=1}^{T}\left(Z_{is}-\widehat{Z}_{is}\right)+\mathbf{X}_{it}'\boldsymbol{\gamma}_{K_0-1}(t/T), & i\in\mathcal{G}_{K_0-1};\\ \epsilon_{it}-\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}+\frac{1}{T}\sum_{s=1}^{T}\left(Z_{is}-\widehat{Z}_{is}\right)+\mathbf{X}_{it}'\boldsymbol{\gamma}_{K_0}(t/T), & i\in\mathcal{G}_{K_0}.\end{cases}$$

Let $\boldsymbol{\Delta}_k=\frac{1}{|\mathcal{G}_k|}\sum_{i\in\mathcal{G}_k}\boldsymbol{\Delta}_i$, $k=1,\ldots,K_0$, and define

$$\boldsymbol{\gamma}_{K_0-1|K_0-1}(u)=\left(|\mathcal{G}_{K_0-1}|\boldsymbol{\Delta}_{K_0-1}+|\mathcal{G}_{K_0}|\boldsymbol{\Delta}_{K_0}\right)^{-1}\left(|\mathcal{G}_{K_0-1}|\boldsymbol{\Delta}_{K_0-1}\boldsymbol{\gamma}_{K_0-1}(u)+|\mathcal{G}_{K_0}|\boldsymbol{\Delta}_{K_0}\boldsymbol{\gamma}_{K_0}(u)\right)$$

as the "pseudo" functional coefficient vector corresponding to $\mathcal{G}_{K_0-1|K_0-1}$, which is essentially a weighted average of $\boldsymbol{\gamma}_{K_0-1}(\cdot)$ and $\boldsymbol{\gamma}_{K_0}(\cdot)$ (note that when $\mathbf{X}_{it}$ are identically distributed over i, then $\boldsymbol{\gamma}_{K_0-1|K_0-1}(u)$ reduces to $\left(|\mathcal{G}_{K_0-1}|+|\mathcal{G}_{K_0}|\right)^{-1}\left(|\mathcal{G}_{K_0-1}\boldsymbol{\gamma}_{K_0-1}(u)+|\mathcal{G}_{K_0}|\boldsymbol{\gamma}_{K_0}(u)\right)$). By (B.22), (B.31) and following the proof of Proposition B.1, we may show that

$$\sup_{\omega\le u\le 1-\omega}\left\|\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(u)-\boldsymbol{\gamma}_{K_0-1|K_0-1}(u)\right\|=O_P\left(h^2+\eta_1(T,h)\right). \tag{B.43}$$

By (B.22), (B.31) and (B.43), we can prove that, uniformly over $i\in\mathcal{G}_{K_0-1}$ and t satisfying $\omega\le t/T\le 1-\omega$,

$$Y_{it}^c-\mathbf{X}_{it}'\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(t/T)=\epsilon_{it}+\mathbf{X}_{it}'\boldsymbol{\nu}_1(t/T)+O_P\left(h^2+\eta_1(T,h)\right),$$

where $\boldsymbol{\nu}_1(u) = \boldsymbol{\gamma}_{K_0-1}(u) - \boldsymbol{\gamma}_{K_0-1|K_0-1}(u)$. Consequently, we have

$$\begin{aligned}
& \frac{1}{NT}\sum_{i\in\mathcal{G}_{K_0-1}}\sum_{t=1}^{T}\left[Y_{it}^c - \mathbf{X}_{it}'\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(t/T)\right]^2 W(t/T) \\
= & \frac{1}{NT}\sum_{i\in\mathcal{G}_{K_0-1}}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T) + \frac{1}{NT}\sum_{i\in\mathcal{G}_{K_0-1}}\sum_{t=1}^{T}\boldsymbol{\nu}_1'(t/T)\mathbf{X}_{it}\mathbf{X}_{it}'\boldsymbol{\nu}_1(t/T)W(t/T) + \\
& \frac{2}{NT}\sum_{i\in\mathcal{G}_{K_0-1}}\sum_{t=1}^{T}\epsilon_{it}\mathbf{X}_{it}'\boldsymbol{\nu}_1(t/T)W(t/T) + o_P(1) \\
\geq & \frac{1}{NT}\sum_{i\in\mathcal{G}_{K_0-1}}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T) + (\underline{\lambda}\cdot\tau_1)\int_{\omega}^{1-\omega}\|\boldsymbol{\nu}_1(u)\|_2^2 W(u)du + o_P(1), \qquad \text{(B.44)}
\end{aligned}$$

where $\underline{\lambda}$ and τ_1 were defined in Assumption 2(iii) and 5(i), respectively. Analogously, we can also prove that

$$\begin{aligned}
& \frac{1}{NT}\sum_{i\in\mathcal{G}_{K_0}}\sum_{t=1}^{T}\left[Y_{it}^c - \mathbf{X}_{it}'\widehat{\boldsymbol{\gamma}}_{K_0-1|K_0-1}(t/T)\right]^2 W(t/T) \\
\geq & \frac{1}{NT}\sum_{i\in\mathcal{G}_{K_0}}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T) + (\underline{\lambda}\cdot\tau_1)\int_{\omega}^{1-\omega}\|\boldsymbol{\nu}_2(u)\|_2^2 W(u)du + o_P(1), \qquad \text{(B.45)}
\end{aligned}$$

where $\boldsymbol{\nu}_2(u) = \boldsymbol{\gamma}_{K_0}(u) - \boldsymbol{\gamma}_{K_0-1|K_0-1}(u)$.

Combining (B.42), (B.44), (B.45) and Assumption 5, we have

$$\begin{aligned}
\mathbb{IC}(K_0-1) &= \log\mathbb{V}_n^2(K_0-1) + (K_0-1)\cdot\rho \\
&\geq \log\left\{\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T) + (\underline{\lambda}\cdot\tau_1)\int_{\omega}^{1-\omega}\left[\|\boldsymbol{\nu}_1(u)\|_2^2 + \|\boldsymbol{\nu}_2(u)\|_2^2\right]W(u)du\right\} + \\
&\quad (K_0-1)\cdot\rho + o_P(1) \\
&\geq \log\left\{\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T) + \frac{1}{2}(\underline{\lambda}\tau_1\tau_2)\right\} + (K_0-1)\cdot\rho + o_P(1) \\
&> \log\left\{\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T)\right\} + K_0\cdot\rho + o_P(1) \\
&= \mathbb{IC}(K_0) + o_P(1). \qquad \text{(B.46)}
\end{aligned}$$

The proof of Lemma B.1 has been completed. □

LEMMA B.2. *Suppose that the assumptions in Theorem 2 are satisfied. Then we have*

$$\mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K), K_0 + 1 \leq K \leq \widetilde{K}\right) \to 1 \tag{B.47}$$

conditional on the event $\mathcal{M}(\mathcal{G})$.

PROOF OF LEMMA B.2. As in the proof of Lemma B.1, without loss of generality, we only consider the case of $K = K_0 + 1$ and prove that

$$\mathbb{P}\left(\mathbb{IC}(K_0) < \mathbb{IC}(K_0 + 1), \mathcal{M}(\mathcal{G})\right) \to 1. \tag{B.48}$$

Conditional on the event $\mathcal{M}(\mathcal{G})$, one of the clusters of $\mathcal{G}_1, \cdots, \mathcal{G}_{K_0}$ are split into two sub-clusters when the HAC algorithm stops at $K = K_0 + 1$. Without loss of generality, we assume that $\mathcal{G}_{K_0}$ is divided into two sub-clusters and denote the resulting $K_0 + 1$ clusters as $\mathcal{G}_1^*, \cdots, \mathcal{G}_{K_0}^*, \mathcal{G}_{K_0+1}^*$ with $\mathcal{G}_k^* = \mathcal{G}_k$ for $k = 1, \cdots, K_0 - 1$ and $\mathcal{G}_{K_0}^* \bigcup \mathcal{G}_{K_0+1}^* = \mathcal{G}_{K_0}$. In this case, the group structure is over-identified.

Observe that

$$\mathbb{V}_n^2(K_0 + 1) = \sum_{k=1}^{K_0+1} \frac{1}{NT} \sum_{i \in \mathcal{G}_k^*} \sum_{t=1}^{T} \left[Y_{it}^c - \mathbf{X}_{it}' \widehat{\boldsymbol{\gamma}}_{k|K_0+1}(t/T)\right]^2 W(t/T) =: \sum_{k=1}^{K_0+1} \mathbb{V}_n^2(k|K_0 + 1).$$

For any $k = 1, \cdots, K_0 + 1$, we write

$$\begin{aligned} Y_{it}^c - \mathbf{X}_{it}' \widehat{\boldsymbol{\gamma}}_{k|K_0+1}(t/T) &= \epsilon_{it} - \mathbf{X}_{it}' \left[\widehat{\boldsymbol{\gamma}}_{k|K_0+1}(t/T) - \boldsymbol{\gamma}_k^*(t/T)\right] - \frac{1}{T}\sum_{s=1}^{T} \epsilon_{is} + \frac{1}{T}\sum_{s=1}^{T}\left(Z_{is} - \widehat{Z}_{is}\right) \\ &=: \epsilon_{it} - \mathbf{X}_{it}' \left[\widehat{\boldsymbol{\gamma}}_{k|K_0+1}(t/T) - \boldsymbol{\gamma}_k^*(t/T)\right] + Q_i, \end{aligned} \tag{B.49}$$

where

$$Q_i = -\frac{1}{T}\sum_{s=1}^{T} \epsilon_{is} + \frac{1}{T}\sum_{s=1}^{T}\left(Z_{is} - \widehat{Z}_{is}\right)$$

and $\boldsymbol{\gamma}_k^*(u) = \boldsymbol{\gamma}_k(u)$ if $k = 1, \cdots, K_0 - 1$ and $\boldsymbol{\gamma}_k^*(u) = \boldsymbol{\gamma}_{K_0}(u)$ if $k = K_0$ and $K_0 + 1$. Note that

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\left(Z_{it} - \widehat{Z}_{it}\right) &= \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{p-1} X_{it,k}\left[\widehat{\beta}_{i,k}(t/T) - \beta_{i,k}(t/T)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\widetilde{\mathbf{X}}_{it}'(\mathbf{0}_{p-1}, \mathbf{I}_{p-1})\left[\widehat{\boldsymbol{\beta}}_i^*(t/T) - \boldsymbol{\beta}_i^*(t/T)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\widetilde{\mathbf{X}}_{it}'(\mathbf{0}_{p-1}, \mathbf{I}_{p-1})\left[\sum_{s=1}^{T}\mathbf{X}_{is}\mathbf{X}_{is}'K_{st}\right]^{-1}\left[\sum_{s=1}^{T}\mathbf{X}_{is}\epsilon_{is}K_{st}\right] + O_P(h^2) \\
&= \frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}\mathbf{X}_{is}'\left[\frac{1}{Th}\sum_{t=1}^{T}\left(\frac{1}{Th}\sum_{s=1}^{T}\mathbf{X}_{is}\mathbf{X}_{is}'K_{st}\right)^{-1}(\mathbf{0}_{p-1}, \mathbf{I}_{p-1})'\widetilde{\mathbf{X}}_{it}K_{st}\right] + O_P(h^2),
\end{aligned}$$

where $\widetilde{\mathbf{X}}_{it} = (X_{it,1}, \cdots, X_{it,p-1})'$, $K_{st} = K\left(\frac{s-t}{Th}\right)$, $\mathbf{0}_k$ is a k-dimensional null vector and $\mathbf{I}_k$ is a $k \times k$ identity matrix. By (B.15), we have

$$\frac{1}{Th}\sum_{s=1}^{T}\mathbf{X}_{is}\mathbf{X}_{is}'K_{st} = \boldsymbol{\Delta}_i + O_P(\eta_1(T,h))$$

uniformly over $1 \le i \le N$ and $1 \le t \le T$, and similarly

$$\frac{1}{Th}\sum_{t=1}^{T}\widetilde{\mathbf{X}}_{it}K_{st} = \widetilde{\boldsymbol{\Gamma}}_i + O_P(\eta_1(T,h))$$

uniformly over $1 \le i \le N$ and $1 \le s \le T$, where $\widetilde{\boldsymbol{\Gamma}}_i = \mathbb{E}(\widetilde{\mathbf{X}}_{it})$. Hence,

$$\frac{1}{T}\sum_{t=1}^{T}\left(Z_{it} - \widehat{Z}_{it}\right) = \frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}\mathbf{X}_{is}'\boldsymbol{\Delta}_i^{-1}(\mathbf{0}_{p-1}, \mathbf{I}_{p-1})'\left(\widetilde{\boldsymbol{\Gamma}}_i + O_P(\eta_1(T,h)\right) + O_P(h^2).$$

All the above implies

$$\begin{aligned}
Q_i =& -\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is} + \frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}\mathbf{X}_{is}'\boldsymbol{\Delta}_i^{-1}(\mathbf{0}_{p-1}, \mathbf{I}_{p-1})'\left(\widetilde{\boldsymbol{\Gamma}}_i + O_P(\eta_1(T,h)\right) + O_P(h^2) \\
=& O_P\left(\eta_2(T) + h^2\right). \qquad \text{(B.50)}
\end{aligned}$$

uniformly over $i = 1, \cdots, N$. The last equality in (B.50) holds because of (B.22) and the similar

result that $\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}\mathbf{X}'_{is} = O_P(\eta_2(T))$ uniformly over $1 \le i \le N$. From (B.50), we readily have

$$\begin{aligned}\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}Q_i^2W(t/T) =& \left(\frac{1}{N}\sum_{i\in\mathcal{G}_k^*}Q_i^2\right)\left(\frac{1}{T}\sum_{t=1}^{T}W(t/T)\right)\\ =&O_P\left(\frac{|\mathcal{G}_k^*|}{N}\left(\frac{\log T}{T}+h^4\right)\right). \end{aligned} \tag{B.51}$$

By (B.22) and (B.50), we may show that

$$\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}Q_i\sum_{t=1}^{T}\epsilon_{it}W(t/T) = O_P\left(\frac{|\mathcal{G}_k^*|}{N}\left(\frac{\log T}{T}+h^4\right)\right). \tag{B.52}$$

By (B.15), (B.50) and the Taylor expansion of $\boldsymbol{\gamma}_k^*(\cdot)$, we readily have

$$\begin{aligned}&\widehat{\boldsymbol{\gamma}}_{k|K_0+1}(u) - \boldsymbol{\gamma}_k^*(u)\\ =&\left[\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}'_{it}K_t(u)\right]^{-1}\left[\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\mathbf{X}_{it}(\epsilon_{it}+Q_i)K_t(u)\right]+O_P(h^2)\\ =&\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}(\boldsymbol{\Delta}_i+O_P(\eta_1(T,h))\right)^{-1}\left(\frac{1}{|\mathcal{G}_k^*|Th}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\mathbf{X}_{it}\epsilon_{it}K_t(u)\right)+\\ &\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}(\boldsymbol{\Delta}_i+O_P(\eta_1(T,h))\right)^{-1}\left(\frac{1}{|\mathcal{G}_k^*|Th}\sum_{i\in\mathcal{G}_k^*}Q_i\sum_{t=1}^{T}\mathbf{X}_{it}K_t(u)\right)+O_P(h^2)\\ =&\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}\boldsymbol{\Delta}_i\right)^{-1}\left(\frac{1}{|\mathcal{G}_k^*|Th}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\mathbf{X}_{it}\epsilon_{it}K_t(u)\right)(1+O_P(\eta_1(T,h)))+\\ &\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}\boldsymbol{\Delta}_i\right)^{-1}\left[\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}\boldsymbol{\Gamma}_i\left(\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}\mathbf{X}'_{is}\right)\boldsymbol{\Delta}_i^{-1}(\mathbf{0}_{p-1},\mathbf{I}_{p-1})'\widetilde{\boldsymbol{\Gamma}}_i\right](1+O_P(\eta_1(T,h)))-\\ &\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}\boldsymbol{\Delta}_i\right)^{-1}\left[\frac{1}{|\mathcal{G}_k^*|}\sum_{i\in\mathcal{G}_k^*}\boldsymbol{\Gamma}_i\left(\frac{1}{T}\sum_{s=1}^{T}\epsilon_{is}\right)\right](1+O_P(\eta_1(T,h)))+O_P(h^2)\\ =:&\mathbf{R}_{k,1}(u)+\mathbf{R}_{k,2}(u)+\mathbf{R}_{k,3}(u)+O_P(h^2)\end{aligned} \tag{B.53}$$

uniformly for $u \in [h, 1-h]$, where $K_t(u) = K\big(\frac{t-uT}{Th}\big)$ as in the proof of Proposition B.1 and $\boldsymbol{\Gamma}_i = \mathbb{E}(\mathbf{X}_{it})$.

Following arguments similar to the proofs of (B.51) and (B.52), we may show that

$$\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\left\{\mathbf{X}_{it}'\left[\mathbf{R}_{k,2}(t/T)+\mathbf{R}_{k,3}(t/T)+O_P(h^2)\right]\right\}^2 W(t/T)=O_P\left(\frac{|\mathcal{G}_k^*|}{N}\left(\frac{\log T}{T}+h^4\right)\right) \quad \text{(B.54)}$$

and

$$\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\epsilon_{it}\left\{\mathbf{X}_{it}'\left[\mathbf{R}_{k,2}(t/T)+\mathbf{R}_{k,3}(t/T)+O_P(h^2)\right]\right\} W(t/T)=O_P\left(\frac{|\mathcal{G}_k^*|}{N}\left(\frac{\log T}{T}+h^4\right)\right). \quad \text{(B.55)}$$

By (A.2) in Assumption 2(iii) and (A.4) in Assumption 6, we may show that

$$\begin{aligned}
&\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\left[\mathbf{X}_{it}'\mathbf{R}_{k,1}(t/T)\right]^2 W(t/T)\\
\leq\ &\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\left[\mathbf{X}_{it}'\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{j\in\mathcal{G}_k^*}\boldsymbol{\Delta}_j\right)^{-1}\left(\frac{1}{|\mathcal{G}_k^*|Th}\sum_{j\in\mathcal{G}_k^*}\sum_{s=1}^{T}\mathbf{X}_{js}\epsilon_{js}K_{st}\right)\right]^2 W(t/T)\\
\leq\ &\underline{\lambda}^{-2}\cdot\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\|\mathbf{X}_{it}\|_2^2\left\|\frac{1}{|\mathcal{G}_k^*|Th}\sum_{j\in\mathcal{G}_k^*}\sum_{s=1}^{T}\mathbf{X}_{js}\epsilon_{js}K_{st}\right\|_2^2 W(t/T)\\
=\ &O_P\left(\frac{|\mathcal{G}_k^*|^2T^2h}{N|\mathcal{G}_k^*|^2T^3h^2}\right)=O_P\left(\frac{1}{NTh}\right). \quad \text{(B.56)}
\end{aligned}$$

By (A.5) in Assumption 6, we can prove that

$$\begin{aligned}
&\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\epsilon_{it}\mathbf{X}_{it}'\mathbf{R}_{k,1}(t/T)W(t/T)\\
=\ &\frac{1}{NT}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\epsilon_{it}\mathbf{X}_{it}'\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{j\in\mathcal{G}_k^*}\boldsymbol{\Delta}_j\right)^{-1}\left(\frac{1}{|\mathcal{G}_k^*|Th}\sum_{j\in\mathcal{G}_k^*}\sum_{s=1}^{T}\mathbf{X}_{js}\epsilon_{js}K_{st}\right)W(t/T)\\
=\ &\frac{1}{N|\mathcal{G}_k^*|T^2h}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\sum_{j\in\mathcal{G}_k^*}\sum_{s=1}^{T}\epsilon_{it}\epsilon_{js}K_{st}\mathbf{X}_{it}'\left(\frac{1}{|\mathcal{G}_k^*|}\sum_{j\in\mathcal{G}_k^*}\boldsymbol{\Delta}_j\right)^{-1}\mathbf{X}_{js}W(t/T)\\
=\ &O_P\left(\frac{|\mathcal{G}_k^*|T}{N|\mathcal{G}_k^*|T^2h}\right)=O_P\left(\frac{1}{NTh}\right). \quad \text{(B.57)}
\end{aligned}$$

By (B.49), (B.51), (B.52), (B.54)–(B.57) and Assumption 5(ii), we can prove that

$$\begin{aligned}
\mathbb{IC}(K_0+1) &= \log \mathbb{V}_n^2(K_0+1) + (K_0+1)\cdot\rho \\
&= \log\left\{\frac{1}{NT}\sum_{k=1}^{K_0+1}\sum_{i\in\mathcal{G}_k^*}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T)\right\} + (K_0+1)\cdot\rho + o_P(\rho) \\
&= \log\left\{\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T)\right\} + (K_0+1)\cdot\rho + o_P(\rho) \\
&> \log\left\{\frac{1}{NT}\sum_{k=1}^{K_0}\sum_{i\in\mathcal{G}_k}\sum_{t=1}^{T}\epsilon_{it}^2 W(t/T)\right\} + K_0\cdot\rho + o_P(\rho) \\
&= \mathbb{IC}(K_0) + o_P(1). \qquad \text{(B.58)}
\end{aligned}$$

The proof of Lemma B.2 has been completed. □

References

Arellano, M. (2003). *Panel Data Econometrics*. Oxford University Press, Oxford.

Boneva, L., Linton, O. and Vogt, M. (2015). A semiparametric model for heterogeneous panel data with fixed effects. *Journal of Econometrics*, 188, 327–345.

Bosq, D. (1998). *Nonparametric Statistics for Stochastic Processes: Estimation and Prediction* (2nd Edition). Lecture Notes in Statistics 110, Springer-Verlag, Berlin.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136, 163–188.

Chen, J., Gao, J. and Li, D. (2012). Semiparametric trending panel data models with cross-sectional dependence. *Journal of Econometrics*, 171, 71–85.

Chen, B. and Huang L. (2017). Nonparametric testing for smooth structural changes in panel data models. *Journal of Econometrics*, 202, 245–267.

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

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning* (2nd Edition). Springer, New York.

Hsiao, C. (2014). *Analysis of Panel Data*. Cambridge University Press.

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

Li, D., Chen, J. and Gao, J. (2011). Non-parametric time-varying coefficient panel data models with fixed effects. *The Econometrics Journal*, 14, 387–408.

Li, D., Wei, L. and Zhang, W. (2018). Nonparametric homogeneity pursuit in functional-coefficient models. *Manuscript*.

Robinson, P. M. (1989). Nonparametric estimation of time-varying parameters. *Statistical Analysis and Forecasting of Economic Structural Change* (ed: P. Hackl). Springer, Berlin, pp. 164–253.

Robinson, P. M. (2012). Nonparametric trending regression with cross-sectional dependence. *Journal of Econometrics*, 169, 4–14.

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

Su, L. and Ullah, A. (2011). Nonparametric and semiparametric panel econometric models: estimation and testing. *Handbook of Empirical Economics and Finance* (eds: A. Ullah and D. E. A. Giles). Taylor & Francis Group, New York, pp. 455–497.

Su, L., Wang, X. and Jin, S. (2018). Sieve estimation of time-varying panel data models with latent structures. Forthcoming in *Journal of Business and Economic Statistics*.

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.

Vogt, M. and Linton, O. (2018). Multiscale clustering of nonparametric regression curves. Cemmap working paper CWP08/18.

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

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58, 236–244.

Zhang, Y., Su, L. and Phillips, P. C. B. (2012). Testing for common trends in semiparametric panel data models with fixed effects. *The Econometrics Journal*, 15, 56–100.