Homogeneity Pursuit in Single Index Models based Panel Data Analysis

Heng Lian
Department of Mathematics
City University of Hong Kong, Kowloon, Hong Kong

Xinghao Qiao
Department of Statistics
London School of Economics, United Kingdom

Wenyang Zhang *
Department of Mathematics
The University of York, United Kingdom

Abstract

Panel data analysis is an important topic in statistics and econometrics. Traditionally, in panel data analysis, all individuals are assumed to share the same unknown parameters, e.g. the same coefficients of covariates when the linear models are used, and the differences between the individuals are accounted for by cluster effects. This kind of modelling only makes sense if our main interest is on the global trend, this is because it would not be able to tell us anything about the individual attributes which are sometimes very important. In this paper, we propose a modelling based on the single index models embedded with homogeneity for panel data analysis, which builds the individual attributes in the model and is parsimonious at the same time. We develop a data driven approach to identify the structure of homogeneity, and estimate the unknown parameters and functions based on the identified structure. Asymptotic properties of the resulting estimators are established. Intensive simulation studies conducted in this paper also show the resulting estimators work very well when sample size is finite. Finally, the proposed modelling is applied to a public financial dataset and a UK climate dataset, the results reveal some interesting findings.

*The corresponding author, email: wenyang.zhang@york.ac.uk.

Keywords and phrases: *Binary segmentation, B-Spline, homogeneity pursuit, single index models.*

1 Introduction

1.1 Preamble

Panel data analysis is an important topic in statistics and econometrics. The traditional approach for analysing panel data assumes all individuals share the same unknown parameters, and uses cluster effects to account for the difference between individuals. For example, when the linear models are used, the coefficients of the covariates are assumed to be the same across all individuals, i.e.

$$y_{it} = \mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_{it}, \quad i = 1, \cdots, m; \quad t = 1, \cdots, T,$$

where y_{it} and $\mathbf{X}_{it} = (X_{it,0}, \ldots, X_{it,p})^{\mathrm{T}}$ with $X_{it,0} = 1$, a $(p+1)$-dimensional vector, are respectively the tth observations of the response variable and covariates of the ith individual. ϵ_{it}, $t = 1, \ldots, T$, are correlated for any given i, and the cluster effects are included in ϵ_{it}. See Hsiao (2014) and the references therein. Whilst this modelling idea is useful when the global trend of the impact of a covariate on the response variable is of our main interest, it does not tell us anything about the individual attributes which are sometimes very important.

In order to explore the individual attributes, we need to make them more concrete and distinctive in modelling. A simple approach to do so would be using

$$y_{it} = \mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i + \epsilon_{it}, \quad i = 1, \cdots, m; \quad t = 1, \cdots, T, \tag{1.1}$$

to fit the data. However, this modelling approach would result in $m(p+1)$ unknown coefficients to estimate, which is too many, because m is usually of the magnitude of hundreds, or even more, in practice. This modelling also ignores the similarity which may exist among some individuals. Such similarity may have very important practical meaning, and could lead to some important findings in practice. In addition to that, statistically speaking, the modelling, like (1.1) without any conditions imposed, would also pay a price on variance side of the estimators resulted because the available information is not used up.

In order to explore the individual attributes and account for the similarity among some individuals at the same time, Ke et al. (2015) proposed a penalised likelihood/least squares based approach to pursue the homogeneity in the linear models, i.e. (1.1), used for panel data analysis, under the framework of treating homogeneity as a kind of sparsity. Regression under homogeneity condition has also been studied by quite a few recent works, e.g. Tibshirani et al. (2005); Friedman et al. (2007); Bondell and Reich (2008); Jiang et al. (2013) , and the references therein. Like Ke et al. (2015), the methods in these works are all based on penalised likelihood/least squares. Ke et al. (2016) took a different approach, they formulated the homogeneity pursuit

problem as a problem of change point detection and applied the binary segmentation approach to identify the homogeneity in the linear models with interactive effects.

The existing literature about homogeneity pursuit mainly focuses on the linear models. It is well known that the linearity condition may not hold for many datasets, and the exploration of linear relationship is not sufficient in many cases. As a consequence, the semiparametric modelling is becoming more and more useful in panel data analysis. Among various semiparametric models, the single index models have many advantages, and are a very successful tool in data analysis, see Härdle and Stoker (1989); Carroll et al. (1997); Yu and Ruppert (2002); Zhu and Xue (2006); Xia (2008); Peng and Huang (2011); Zhu et al. (2012); Guo et al. (2017), and the reference therein. In this paper, we are going to investigate the homogeneity pursuit in the single index models used for panel data analysis. The detailed definition of the models we are going to address in this paper is given in Section 1.2

1.2 The single index models with homogeneity structure

Let y_{it} and $\mathbf{X}_{it}$, a $(p+1)$-dimensional vector, be respectively the tth observations of the response variable and covariate of the ith individual, $i = 1, \cdots, m$; $t = 1, \cdots, T$. We consider the models

$$y_{it} = g_i(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) + \epsilon_{it}, \quad i = 1, \cdots, m; \quad t = 1, \cdots, T, \tag{1.2}$$

where

$$g_i(\cdot) = \begin{cases} g_{(1)}(\cdot) & \text{when } i \in G_{1,1}, \\ g_{(2)}(\cdot) & \text{when } i \in G_{1,2}, \\ \vdots & \vdots \\ g_{(H_1)}(\cdot) & \text{when } i \in G_{1,H_1}, \end{cases} \qquad \beta_{ij} = \begin{cases} \beta_{(1)} & \text{when } (i,j) \in G_{2,1}, \\ \beta_{(2)} & \text{when } (i,j) \in G_{2,2}, \\ \vdots & \vdots \\ \beta_{(H_2)} & \text{when } (i,j) \in G_{2,H_2}, \end{cases} \tag{1.3}$$

$\mathbb{G}_1 = \{G_{1,k} : k = 1, \cdots, H_1\}$ is a partition of set $\{1, \cdots, m\}$, $\mathbb{G}_2 = \{G_{2,k} : k = 1, \cdots, H_2\}$ is a partition of set $\{(i,j) : i = 1, \cdots, m; j = 1, \cdots, p\}$, β_{ij} is the $(j+1)$th component of $\boldsymbol{\beta}_i$, and $E(\epsilon_{it}|\mathbf{X}_{it}) = 0$. The condition (1.3) is the homogeneity structure of the standard single index models for panel data analysis. $\{G_{1,k} : k = 1, \cdots, H_1\}$ and $\{G_{2,k} : k = 1, \cdots, H_2\}$ are unknown partitions. H_1 and H_2 are unknown integers, H_1 is much smaller than m, H_2 is much smaller than mp. $g_{(k)}(\cdot)$, $k = 1, \cdots, H_1$, are unknown functions to be estimated, and $\beta_{(k)}$, $k = 1, \cdots, H_2$, are unknown parameters to be estimated.

As a special case of the proposed models, if $g_i(\cdot)$s are homogeneous but coefficients are heterogeneous, the models can be simplified to

$$y_{it} = g(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) + \epsilon_{it}, \quad i = 1, \cdots, m; \quad t = 1, \cdots, T,$$

with

$$\beta_{ij} = \begin{cases} \beta_{(1)} & \text{when } (i,j) \in G_{2,1}, \\ \beta_{(2)} & \text{when } (i,j) \in G_{2,2}, \\ \vdots & \vdots \\ \beta_{(H_2)} & \text{when } (i,j) \in G_{2,H_2}. \end{cases}$$

Let $\beta_{i0} = 1$ be the first component of $\boldsymbol{\beta}_i$. In the literature, the most commonly used identification condition for the single index models is $\|\boldsymbol{\beta}_i\| = 1$ and $\beta_{i0} > 0$, or $\beta_{i0} = 1$. We choose the latter in this paper.

From model specification point of view, it is easy to see the commonly used fixed effect models are a special case of (1.2). In fact, when $g_i(\cdot) = g(\cdot) + \alpha_i$ and $\boldsymbol{\beta}_i = \boldsymbol{\beta}$ in (1.2), (1.2) becomes

$$y_{it} = g(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}) + \alpha_i + \epsilon_{it},$$

which are the standard single index fixed effect models.

The models (1.2) together with (1.3) show that the homogeneity pursuit in the single index models for panel data analysis is even more important than that in the linear models, this is because we would have to estimate m unknown functions and mp unknown parameters in order to explore the individual attributes, if the homogeneity pursuit is not conducted. However, if the homogeneity pursuit is conducted, we only need to estimate H_1, much smaller than m, unknown functions and H_2, much smaller than mp, unknown parameters when the homogeneity exists. Even without taking into account the benefit resulted from the homogeneity pursuit for the parametric part of the models, just for the part of unknown functions alone, to estimate much fewer functions would make a big difference in the obtained estimators, in terms of the stability of the estimators.

In addition to the methodological advantage, the proposed homogeneity pursuit also acts as a detector to find which covariates have the same impact on the response variable. For example, when the response variable is temperature and the linear models are employed, if the homogeneity pursuit finds the coefficients of two environmental factors in the models are the same, that would mean these two environmental factors have the same impact on temperature, which would be practically very meaningful.

There is an expanding literature on panel data models with grouping structure. Ke et al. (2016) used a similar binary segmentation method to identify grouping structure in panel data models with interactive effects. Applying the idea of penalised least squares based CARDS, proposed by Ke et al. (2015), Wang et al. (2018) designed an algorithm called Panel-CARDS to identify group structure in linear models for panel data analysis. Wang and Su (2019) considered very general parametric nonlinear models using binary segmentation. The grouping structures in these models are determined by finite-dimensional parameters. As the models in Wang and Su (2019) are defined through a general known function without any specific form, the theoretical results apply for a large class of parametric nonlinear models. For nonparametric

models, Vogt and Linton (2017) used k-means algorithm to cluster the individual functions. In this paper, we adopt the binary segmentation approach for grouping structure identification. The unique feature and our main contribution is that we consider semiparametric single-index models in which the grouping structure is determined by both nonparametric functions and finite-dimensional index parameters.

Like any statistical modelling, the proposed models have its own limitations. One limitation is that we assume exact homogeneity within a group (parameters within the same group are exactly the same). Such an assumption was used in all references above. One exception is Bonhomme et al. (2017) which allows the parameters to be approximately equal within a group, with an additional term characterizing the approximation error in their theoretical results. Similar approximation error terms also appear in Belloni et al. (2014) for a different study. An extension to allow for group-specific effects of covariates is briefly discussed in Bonhomme and Manresa (2015). We conjecture that assumptions and proofs in our modelling framework can also be modified to allow such cases.

The rest of the paper is organized as follows. We begin in Section 2 with a description of the proposed estimation procedure which is embedded with a binary segmentation based homogeneity pursuit. The asymptotic properties of the proposed estimators are presented in Section 3. The finite-sample performance of the proposed estimation, in comparison with the potential competitors, is assessed by simulation studies in Section 4. In Section 5, applying the single index models (1.2) together with the homogeneity structure (1.3) to the 49 Industry Portfolios data set, which can be freely downloaded from Kenneth French's website

`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`,

and the UK climate data, which can be freely downloaded from

`http://www.metoffice.gov.uk/public/weather/climate-historic`,

we will show the advantages of the proposed statistical methodology. We leave all technical proofs of the asymptotic properties in the Appendix and the online supplementary material.

2 Estimation procedure

2.1 Estimation method

Our approach to deal with the unknown functions $g_i(\cdot)$, $i = 1, \cdots, m$, in (1.2) is based on the B-Spline. To achieve the best result for the homogeneity pursuit, we have to decompose all $g_i(\cdot)$s by the same B-Spline basis, $\mathbf{B}(\cdot) = (B_1(\cdot), \cdots, B_K(\cdot))^{\mathrm{T}}$. B-splines and kernel smoothing are two major approaches to estimate unknown functions in statistics literature. Either has its advantage and disadvantage. Kernel smoothing demands computation of estimated function value on a grid of points, which can be computationally very expensive, we therefore go for B-

splines in this paper. Another advantage of using B-splines is it makes the binary segmentation based homogeneity pursuit much easier.

For each i, $i = 1, \cdots, m$, let $\tilde{\boldsymbol{\beta}}_i$ be the estimate of $\boldsymbol{\beta}_i$ obtained, based on the observations for the ith individual, by a standard estimation procedure for the single index models, e.g. the method in Yu and Ruppert (2002) or in Härdle and Stoker (1989), and

$$a = \min_{1\le i\le m} \min_{1\le t\le T} \mathbf{X}_{it}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_i, \quad b = \max_{1\le i\le m} \max_{1\le t\le T} \mathbf{X}_{it}^{\mathrm{T}}\tilde{\boldsymbol{\beta}}_i.$$

We use the B-Spline basis of order s in this paper, and the basis, $\mathbf{B}(\cdot)$, is formed by the equally spaced knots, τ_k, $k = 0, \cdots, K - s + 1$, on the interval $[a, b]$, with $\tau_0 = a$ and $\tau_{K-s+1} = b$. Based on the basis $\mathbf{B}(\cdot)$, $g_i(\cdot)$ can be decomposed as

$$g_i(\cdot) \approx \mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\theta}_i, \tag{2.1}$$

where $\boldsymbol{\theta}_i = (\theta_{i1}, \cdots, \theta_{iK})^{\mathrm{T}}$. So, to get the estimator of $g_i(\cdot)$, we only need to get the estimator of $\boldsymbol{\theta}_i$. Theoretically, we have to assume there exists a $\boldsymbol{\theta}_{0i}$ such that $\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\theta}_{0i}$ is a good approximation of $g_i(\cdot)$. Instead of making this assumption directly, we will assume that $g_i(\cdot)$ is twice differentiable, which implies that as long as $s > 2$ we can find $\boldsymbol{\theta}_{0i}$ such that $|g_i(x) - \mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\theta}_{0i}| \le CK^{-2}$ for some constant $C > 0$. We will also assume $K \to \infty$ and thus the approximation error converges to zero.

Our estimation procedure for $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$, $i = 1, \cdots, m$, consists of three stages: in the first stage, for each i, we estimate $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}_i$ only based on the observations for the ith individual, and treat the obtained estimators as initial estimators; we identify, in the second stage, the homogeneity structure in the $\boldsymbol{\theta}_i$s and $\boldsymbol{\beta}_i$s based on the initial estimators obtained in the first stage; in the final stage, we estimate the $\boldsymbol{\theta}_i$s and $\boldsymbol{\beta}_i$s under the identified homogeneity structure.

We now present the details of the estimation procedure.

Stage 1 (*Initial Estimation*). Let $\overline{\boldsymbol{\beta}}_i = (\beta_{i1}, \cdots, \beta_{ip})^{\mathrm{T}}$, which is $\boldsymbol{\beta}_i$ with the first component, which is always 1, being dropped. For each i, based on the observations for the ith individual, approximating $g_i(\cdot)$ by its decomposition (2.1) and applying the least squares estimation method, we have the following objective function

$$\sum_{t=1}^{T} \Big(y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i)\boldsymbol{\theta}_i\Big)^2. \tag{2.2}$$

Minimise (2.2) with respect to $(\overline{\boldsymbol{\beta}}_i^{\mathrm{T}}, \boldsymbol{\theta}_i^{\mathrm{T}})$, and denote the resulting minimiser by $(\tilde{\boldsymbol{\beta}}_i^{\mathrm{T}}, \tilde{\boldsymbol{\theta}}_i^{\mathrm{T}})$. We will show how to conduct the minimisation in Section 2.2.

Stage 2 (*Homogeneity Pursuit*). Let $\tilde{\beta}_{ij}$ be the jth component of $\tilde{\boldsymbol{\beta}}_i$, we sort $\tilde{\beta}_{ij}$, $i = 1, \cdots, m$, $j = 1, \cdots, p$, in ascending order, and denote them by

$$b_{(1)} \le \cdots \le b_{(mp)}.$$

We use R_{ij} to denote the rank of $\tilde{\beta}_{ij}$. Identifying the homogeneity among $\tilde{\beta}_{ij}$, $i = 1, \cdots, m$, $j = 1, \cdots, p$, is equivalent to detecting the change points among $b_{(l)}$, $l = 1, \cdots, mp$. To this end, we apply the Binary Segmentation algorithm as follows.

For any $1 \le i < j \le mp$, let

$$\Delta_{ij}(\kappa) = \sqrt{\frac{(j-\kappa)(\kappa-i+1)}{j-i+1}} \left| \frac{\sum_{l=\kappa+1}^{j} b_{(l)}}{j-\kappa} - \frac{\sum_{l=i}^{\kappa} b_{(l)}}{\kappa-i+1} \right|.$$

Given a threshold δ, the Binary Segmentation algorithm to detect the change points works as follows

(1) Find $\hat{k}_1$ such that

$$\Delta_{1,mp}(\hat{k}_1) = \max_{1 \le \kappa < mp} \Delta_{1,mp}(\kappa).$$

If $\Delta_{1,mp}(\hat{k}_1) \le \delta$, there is no change point among $b_{(l)}$, $l = 1, \cdots, mp$, and the process of detection ends. Otherwise, add $\hat{k}_1$ to the set of change points and divide the region $\{\kappa : 1 \le \kappa \le mp\}$ into two subregions: $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$ and $\{\kappa : \hat{k}_1 + 1 \le \kappa \le mp\}$.

(2) Detect the change points in the two subregions obtained in (1), respectively. Let us deal with the region $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$ first. Find $\hat{k}_2$ such that

$$\Delta_{1,\hat{k}_1}(\hat{k}_2) = \max_{1 \le \kappa < \hat{k}_1} \Delta_{1,\hat{k}_1}(\kappa).$$

If $\Delta_{1,\hat{k}_1}(\hat{k}_2) \le \delta$, there is no change point in the region $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$. Otherwise, add $\hat{k}_2$ to the set of change points and divide the region $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$ into two subregions: $\{\kappa : 1 \le \kappa \le \hat{k}_2\}$ and $\{\kappa : \hat{k}_2 + 1 \le \kappa \le \hat{k}_1\}$. For the region $\{\kappa : \hat{k}_1 + 1 \le \kappa \le mp\}$, we find $\hat{k}_3$ such that

$$\Delta_{\hat{k}_1+1,mp}(\hat{k}_3) = \max_{\hat{k}_1+1 \le \kappa < mp} \Delta_{\hat{k}_1+1,mp}(\kappa).$$

If $\Delta_{\hat{k}_1+1,mp}(\hat{k}_3) \le \delta$, there is no change point in the region $\{\kappa : \hat{k}_1 + 1 \le \kappa \le mp\}$. Otherwise, add $\hat{k}_3$ to the set of change points and divide the region $\{\kappa : \hat{k}_1 + 1 \le \kappa \le mp\}$ into two subregions: $\{\kappa : \hat{k}_1 + 1 \le \kappa \le \hat{k}_3\}$ and $\{\kappa : \hat{k}_3 + 1 \le \kappa \le mp\}$.

(3) For each subregion obtained in (2), we do exactly the same as that for the subregion $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$ or $\{\kappa : \hat{k}_1 + 1 \le \kappa \le mp\}$ in (2), and keep doing so until there is no subregion containing any change point.

We sort the estimated change point locations in ascending order and denote them by

$$\hat{k}_{(1)} < \hat{k}_{(2)} < \cdots < \hat{k}_{(\hat{H}_{-1})},$$

where $\hat{H}_{-1}$ is the number of change points detected. In addition, we denote $\hat{k}_{(0)} = 0$, $\hat{H}_2 = \hat{H}_{-1} + 1$, and $\hat{k}_{(\hat{H}_2)} = mp$.

We use $\hat{H}_2$ to estimate H_2. Let

$$\hat{G}_{2,h} = \{(i,j) : \hat{k}_{(h-1)} < R_{ij} \le \hat{k}_{(h)}\}, \quad 1 \le h \le \hat{H}_2,$$

we use $\left\{\hat{G}_{2,h} : 1 \le h \le \hat{H}_2\right\}$ to estimate the partition $\{G_{2,h} : 1 \le h \le H_2\}$. We assume that all the β_{ij}s with the subscript (i,j) in the same estimated partition have the same value.

Let $\tilde{\theta}_{ij}$ be the jth component of $\tilde{\boldsymbol{\theta}}_i$. Doing exactly the same to $\tilde{\theta}_{ij}$, $i = 1, \cdots, m$, $j = 1, \cdots, K$, we get a partition $\{\hat{G}_{1,1}, \cdots, \hat{G}_{1,\hat{H}_1}\}$ of $\{(i,j) : i = 1, \cdots, m; j = 1, \cdots, K\}$. We assume that all the θ_{ij}s with subscript (i,j) in the same estimated partition have the same value.

Stage 3 (*Final Estimation*). Let $L(\eta_1, \cdots, \eta_{\hat{H}_2}, \xi_1, \cdots, \xi_{\hat{H}_1})$ be

$$\sum_{i=1}^{m}\sum_{t=1}^{T}\Big(y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i)\boldsymbol{\theta}_i\Big)^2. \tag{2.3}$$

with β_{ij}, $i = 1, \cdots, m$, $j = 1, \cdots, p$, being replaced by η_k if $(i,j) \in \hat{G}_{2,k}$, and θ_{ij}, $i = 1, \cdots, m$, $j = 1, \cdots, K$, being replaced by ξ_h if $(i,j) \in \hat{G}_{1,h}$. Let $(\hat{\eta}_1, \cdots, \hat{\eta}_{\hat{H}_2}, \hat{\xi}_1, \cdots, \hat{\xi}_{\hat{H}_1})$ minimise $L(\eta_1, \cdots, \eta_{\hat{H}_2}, \xi_1, \cdots, \xi_{\hat{H}_1})$. The final estimator $\hat{\beta}_{ij}$ of β_{ij} is $\hat{\eta}_k$ if $(i,j) \in \hat{G}_{2,k}$, and the final estimator $\hat{\theta}_{ij}$ of θ_{ij} is $\hat{\xi}_h$ if $(i,j) \in \hat{G}_{1,h}$. Once we have the estimator $\hat{\theta}_{ij}$, the estimator $\hat{g}_i(\cdot)$ of $g_i(\cdot)$ is taken to be $\mathbf{B}(\cdot)^{\mathrm{T}}\hat{\boldsymbol{\theta}}_i$.

Remark 1 *When dealing with the unknown functions $g_i(\cdot)$, $i = 1, \cdots, m$, in the estimation procedure, instead of treating each unknown function as a single undivided unit to conduct homogeneity pursuit, we work on the coefficients of its B-Spline decomposition. This is because there may still be some kind of homogeneity between two functions even if they are different. For example, for two different functions, it could be the case that some coefficients of the B-Spline decomposition of one function are the same as some coefficients of the B-Spline decomposition of another one. If we treat each unknown function as a single undivided unit to conduct homogeneity pursuit, we would not be able to identify or use this kind of homogeneity, which would make our final estimators not as efficient as they should.*

Our estimation procedure for grouping functions $g_i(\cdot)$ is based on estimated values of $\boldsymbol{\theta}_{0i}$. Thus theoretically correct grouping indeed depend on $\boldsymbol{\theta}_{0i}$ which has different dimensions with different choices of K. Theoretically, this does not cause any problem since $g_i(\cdot) = g_j(\cdot)$ implies $\boldsymbol{\theta}_{0i} = \boldsymbol{\theta}_{0j}$, so we can still get a consistent estimator of the partition even for different values of K. As discussed in Section 2.3 below, we fix a reasonable value of K for ease of implementation.

2.2 Computational algorithm

In the estimation procedure described in Section 2.1, the minimiser of (2.2) does not have a closed form, neither does the minimiser of $L(\eta_1, \cdots, \eta_{\hat{H}_2}, \xi_1, \cdots, \xi_{\hat{H}_1})$. To conduct the minimisation of either of the two objective functions, we appeal to the standard NLS algorithm, and use the `nlsLM of minpack.lm` package in R to implement it. One can also use other NLS software, for example, the NLS routine `lsqnonlin()` from MATLAB and `PROC NLIN` from SAS. To use the `nlsLM of minpack.lm` package in R, we first need to find an initial value. The initial value for minimising (2.2) can be obtained as follows:

(1) Apply the standard least squares estimation for the linear models to $(y_{it},\ \mathbf{X}_{it})$, $t = 1,\ \cdots,\ T$, and denote the resulting estimator by $\check{\boldsymbol{\beta}}_i$, the initial value for $\boldsymbol{\beta}_i$ is taken to be $\boldsymbol{\beta}_i^{(0)} = \check{\beta}_{i0}^{-1}\check{\boldsymbol{\beta}}_i$, $\check{\beta}_{i0}$ is the first component of $\check{\boldsymbol{\beta}}_i$.

(2) Substitute $\boldsymbol{\beta}_i^{(0)}$ for $\boldsymbol{\beta}_i$ in (2.2), then minimise (2.2) with respect to $\boldsymbol{\theta}_i$, the minimiser $\boldsymbol{\theta}_i^{(0)}$ is the initial value of $\boldsymbol{\theta}_i$.

Once we have $\boldsymbol{\beta}_i^{(0)}$ and $\boldsymbol{\theta}_i^{(0)}$, the minimiser of (2.2) can be obtained by the `nlsLM of minpack.lm` package in R straightforwardly. We note that using the standard least squares estimation is actually a common way to get initial values for single-index models and this method is used in Carroll et al. (1997). One could also justify this choice based on sufficient dimension reduction (Duan and Li; 1991) which established that even for semiparametric models, linear regression can be used to estimate the index parameter under mild assumptions. In our simulations, this initial value works well. One could also try generating multiple random initial values and fit the model multiple times as a safeguard.

For any set A, let $|A|$ be the number of elements in A. The initial value for minimising $L(\eta_1,\ \cdots,\ \eta_{\hat{H}_2},\ \xi_1,\ \cdots,\ \xi_{\hat{H}_1})$ can be obtained through the initial estimates of $\boldsymbol{\beta}_i$ and $\boldsymbol{\theta}_i$, obtained in Stage 1 of the estimation procedure in Section 2.1, as follows:

$$\eta_s^{(0)} = \left(|\hat{G}_{2,s}|\right)^{-1} \sum_{(i,j)\in\hat{G}_{2,s}} \tilde{\beta}_{ij}, \quad s = 1,\ \cdots,\ \hat{H}_2$$

and

$$\xi_s^{(0)} = \left(|\hat{G}_{1,s}|\right)^{-1} \sum_{(i,j)\in\hat{G}_{1,s}} \tilde{\theta}_{ij}, \quad s = 1,\ \cdots,\ \hat{H}_1.$$

Once we have the initial value $(\eta_1^{(0)},\ \cdots,\ \eta_{\hat{H}_2}^{(0)},\ \xi_1^{(0)},\ \cdots,\ \xi_{\hat{H}_1}^{(0)})$, we can have the minimiser of $L(\eta_1,\ \cdots,\ \eta_{\hat{H}_2},\ \xi_1,\ \cdots,\ \xi_{\hat{H}_1})$ by using the `nlsLM of minpack.lm` package in R straightforwardly.

2.3 Selection of tuning parameters

In literature, s, the order of B-Spline, is almost always fixed to be either $s = 3$ (quadratic splines) or $s = 4$ (cubic splines). In this paper we use the cubic splines. Furthermore, to ease the computational burden, we fix $K = 6$. One could select K using cross-validation or some information criterion but it would increase the computational burden with no appreciable numerical advantages in our experience. For complicated models, using a fixed K is not uncommon, and this is the case, for example, in Huang et al. (2010); Fan et al. (2011). This choice of K is small enough to avoid overfitting in typical problems with sample size not too small, and big enough to flexibly approximate many smooth functions accurately.

The threshold δ in the Stage 2 of the proposed estimation procedure, described in Section 2.1, plays a key role for the success of the homogeneity pursuit. Obviously, when using a decreasing sequence of δ on a fine grid, the number of change points will increase resulting in a nested

sequence of sets of change points. This nested sequence can be identified all at the same time using a small δ. Thus the selection of δ is equivalent to the selection of $\hat{H}_1$ and $\hat{H}_2$ and there is only a finite number of possibilities for the values of the latter. Therefore, in this section, instead of selecting δ, we equivalently describe a cross-validation procedure to select the two tuning parameters, $\hat{H}_1$ and $\hat{H}_2$.

For the single index model (1.2) where $\mathbf{X}_{it}$'s are independent across $t = 1, \cdots, T$, we implement a L-fold cross validation approach. In particular, for a given pair $\{H_1, H_2\}$, we remove $1/L$th of the observed time points for $\{(y_{it}, \mathbf{X}_{it}), i = 1, \cdots, m, t = 1, \cdots, T\}$ as a validation set, estimate the single index model (1.2) with identified homogeneity structure on the remaining data, compute the squared error between y_{it} and fitted values $\hat{g}_i(\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i) = \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i)\hat{\boldsymbol{\theta}}_i$, on the validation set, and repeat this procedure L times to calculate the cross-validated mean squared error and its corresponding standard error. We search over a grid of $\{H_1, H_2\}$ values and apply the one-standard-error rule to choose the smallest model for which the estimated cross-validated error is within one standard error of the lowest point on the error surface. The rationale here is that if a set of models appear to be more or less equally good, then we might tend to choose the simplest model. Across the candidate pairs, $\{\hat{H}_1, \hat{H}_2\}$, whose corresponding errors are within this deviation, one can choose the smallest $\hat{H}_1$ after selecting the smallest $\hat{H}_2$ or switch the selection order or select the smallest value of $\hat{H}_1 + \hat{H}_2$, we take the first approach since it produces better model selection consistency in our numerical experiments. A similar one-standard-deviation-rule technique has been adopted to choose the regularisation parameter with a smaller model size for the lasso problems (James et al.; 2013).

When $\mathbf{X}_{it}$'s are time dependent panel data, we implement a rolling procedure to perform cross-validation for time series. More specifically, for each $r = L, L-1, \cdots, 1$, we rollingly treat $\{(y_{it}, \mathbf{X}_{it}), i = 1, \cdots, m, t = 1, \cdots, T-r\}$ as training observations and $\{(y_{i,T-r+1}, \mathbf{X}_{i,T-r+1}), i = 1, \cdots, m\}$ as validation set, calculate the squared error between each y_{it} and its fitted value. Finally, we apply the one-standard-deviation-rule on the lowest cross-validated mean squared error and choose $\hat{H}_2$ and $\hat{H}_1$.

Cross-validation is often used in statistics and econometrics literature for selection of the tuning parameters. It would be significant to provide theoretical properties for this method. However, this is a challenging problem that we currently do not know how to solve. The model here with the three-stage algorithm is much more complicated than other models for which consistency of cross-validation has been established. We also note that some works (Lu and Su; 2017; Vogt and Matthias; 2017) has used other approaches for determining the number of groups in parametric models.

In the cross-validation procedure when we need to make predictions for the validation set, the domain in $\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i)$ for the training data set might not cover that for the validation set. We adopt the idea in Wang and Yang (2009) by mapping $\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i$ to $F_i(\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i) \sim \mathrm{Unif}[0, 1]$ so that the domain is fixed to be $[0, 1]$, where F_i is the cumulative distribution function of

$\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i$. We then implement the estimation procedure in Section 2.1 by decomposing $g_i(\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i) \approx \mathbf{B}^{\mathrm{T}}\big(\hat{F}_i(\mathbf{X}_{it}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_i)\big)\boldsymbol{\theta}_i$, where $\hat{F}_i$ is the empirical estimate of F_i obtained from $\{\mathbf{X}_{it}\hat{\boldsymbol{\beta}}_i\}_{t=1}^{T}$. The proposed approach is thus able to make predictions and, as demonstrated by some numerical studies, provides very similar sample performance in terms of estimation accuracy.

2.4 Post-processing step

The Binary Segmentation algorithm is known not to perform well in certain unfavourable situations, since it may fit a step function with a single change-point to the data over a segment that possibly contains multiple change-points. Motivated by the post-processing idea in Cho and Fryzlewicz (2012), we equip the Binary Segmentation algorithm with an additional step aiming to enhance the accuracy of detected change-points locations through a fine-scale search. To be specific, at each estimated change-point, we re-calculate $\Delta_{ij}(\kappa)$ over the interval between two adjacent estimated change-points such that each segment only contains a single estimated change-point and then identify the new change-point location to replace the old one. We perform this post-processing procedure by iteratively cycling through all neighbouring change-points and fine-tuning the change-points locations. This procedure is terminated when the set of change-points does not change. Our numerical experiments show that this extra post-processing step apparently improves the accuracy of each estimated change-point location and hence the identified homogeneity structure for model (1.2).

3 Asymptotic properties

In this section, we are going to investigate the asymptotic behaviour of the estimators obtained by the proposed estimation procedure, which we call correct-fitting, and compare with the estimators obtained without homogeneity pursuit, which is the initial estimators obtained in the Stage 1 in the proposed estimation procedure, we call it over-fitting, and the estimators obtained under the assumption that all individuals share the same index (namely, $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_m$), which we call under-fitting. The asymptotic theory presented in this section is in the sense that $T \longrightarrow \infty$, m is possibly diverging to infinity, but p, H_1, H_2 are fixed. This agrees with many applications in which H_1 and H_2 are expected to be small and thus significant reduction of unknown parameters can be achieved by clustering the parameters. To make the presentation neat, we state the asymptotic theorems in this section and leave all technical proofs in the Appendix and the supplementary material. We impose the following assumptions.

(C1) For each i, $(y_{it}, \mathbf{X}_{it}, \epsilon_{it}), t = 1, \ldots, T$ is stationary and α-mixing with mixing coefficient $\alpha_i(l) \leq \rho^l$ for some $\rho \in (0, 1)$, and the m time series are independent.

(C2) ϵ_{it} satisfies $E[\epsilon_{it}|\mathbf{X}_{it}] = 0$ and $E[\epsilon_{it}^d|\mathbf{X}_{it}] < C$ for some constants $C > 0$ and $d > 2$. The variables $X_{it,j}$ are uniformly bounded. The density of $\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}$ is uniformly bounded and

bounded away from zero on its support. The order of the spline satisfies $s > 2$ (s is fixed and does diverge with T).

(C3) The link functions g_{0i} are twice continuously differentiable. We also assume $E[\overline{\mathbf{X}}_{it}|\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i = x]$ is twice continuously differentiable for $\boldsymbol{\beta}_i$ in a neighborhood of $\boldsymbol{\beta}_{0i}$, where $\overline{\mathbf{X}}_{it} = (X_{it,1}, \cdots, X_{it,p})^{\mathrm{T}}$.

(C4) Both $E[\mathbf{X}_{it}\mathbf{X}_{it}^{\mathrm{T}}]$ and $E[(g'_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i))^2(\overline{\mathbf{X}}_{it} - E[\overline{\mathbf{X}}_{it}|\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i])^{\otimes 2}]$ have eigenvalues bounded and bounded away from zero, uniformly over i and $\boldsymbol{\beta}_i$ in a neighborhood of $\boldsymbol{\beta}_{0i}$, where for any matrix $\mathbf{A}$, $\mathbf{A}^{\otimes 2} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$.

(C5) H_1 and H_2 are fixed and $\max_{i,j} m_{ij}/\min_{i,j} m_{ij}$ and $\max_i m_i/\min_i m_i$ are bounded, where m_i is the size of $G_{1,h}$ that contains i, and m_{ij} is the size of $G_{2,h}$ that contains β_{ij}. We set $K \asymp (mT)^{1/5}$. Assume $\frac{m(\log T)}{T} \to 0$.

(C6) Assume $\sqrt{mK\log(Tm)/T} << \delta_1 << \gamma_1$, where γ_1 is the minimum jump size for the sequence $\theta_{0(1)} \leq \cdots \leq \theta_{0(mK)}$ at the change points, and δ_1 is the threshold used in the change point detection algorithm (we stop partitioning if the test statistic is below δ_1). Similarly, assume $\sqrt{mp\log(Tm)/T} << \delta_2 << \gamma_2$, where γ_2 and δ_2 are similarly defined for the sequence $\beta_{0(1)} \leq \cdots \leq \beta_{0(mp)}$.

Remark 2 *(C1) and (C2) contains some mild regularity assumptions. Geometric mixing condition could be relaxed to $\alpha_i(l) \leq l^{-C}$ for sufficiently large C but the expressions in the theoretical derivation would become very messy. (C2) also allows heterogeneous errors. Assuming $X_{it.j}$ to be bounded is common in estimation with B-splines since the basis functions are constructed on a compact interval. In practice, one can always transform the predictors to $[0, 1]$ before analysis although this is not required. (C3) contains smoothness conditions for some functions and (C4) contains some identifiability conditions usually assumed in single-index models and involves the projection one typically use to profile out the nonparametric part. This requires in particular that $g_i(\cdot)$ are non-constant otherwise the index parameter cannot be identified. Uniformity over i in various assumptions above is void if m is fixed. (C5) specifies the required divergence rate for T, m, K. Finally, (C6) is used in showing that stage 2 of our estimation procedure can identify the true partition with probability approaching one. It requires the parameter values in different groups are sufficiently different for the group to be estimated correctly.*

When considering the estimator in stage 1 of our estimation procedure, we can replace (C5) with the following assumption (C5'). For the estimator in stage 1, there is no need to assume H_1, H_2 are fixed or assume all m_{ij} and m_i are of the same size as in (C5).

(C5') We set $K \asymp T^{1/5}$.

We start with the asymptotic properties of the estimators obtained without homogeneity pursuit. The exact expressions of the asymptotic variances in either Theorem 1 or Theorem 2 are too complicated to be presented in the main body of this paper, therefore, left in Appendix.

Theorem 1 *(Over-fitting case). For any i, $i = 1, \cdots, m$, and $1 \le j \le p$, under the conditions (C1)-(C4) and (C5'), we have*

$$T^{1/2}(\widetilde{\mathbf{e}}_{ij}^{\mathrm{T}}\widetilde{\mathbf{\Theta}}_2\widetilde{\mathbf{e}}_{ij})^{-1/2}\left(\tilde{\beta}_{ij}-\beta_{ij}\right)\xrightarrow{D}N(0,\ 1)$$

and

$$T^{2/5}(\widetilde{\mathbf{b}}_i^{\mathrm{T}}(u)\widetilde{\mathbf{\Theta}}_1\widetilde{\mathbf{b}}_i(u))^{-1/2}\left(\widetilde{g}_i(u)-g_i(u)-r_i(u)\right)\xrightarrow{D}N(0,\ 1),$$

where $\widetilde{\mathbf{e}}_{ij}$ and $\widetilde{\mathbf{b}}_i(u)$ are unit vectors, $\widetilde{\mathbf{\Theta}}_1$, $\widetilde{\mathbf{\Theta}}_2$ are matrices with eigenvalues bounded and bounded away from zero, all these quantities are defined in the proof in the Appendix. The bias term $r_i(u) = g_i(u) - \mathbf{B}^{\mathrm{T}}(u)\boldsymbol{\theta}_{0i}$ satisfies $|r_i(u)| \le CK^{-2}$, where $\boldsymbol{\theta}_{0i}$ is the vector of spline coefficients used to approximate g_i as defined in the Appendix.

From Theorem 1, it is easy to see both $\widetilde{\mathbf{e}}_{ij}^{\mathrm{T}}\widetilde{\mathbf{\Theta}}_2\widetilde{\mathbf{e}}_{ij}$ and $\widetilde{\mathbf{b}}_i^{\mathrm{T}}(u)\widetilde{\mathbf{\Theta}}_1\widetilde{\mathbf{b}}_i(u)$ are scalars bounded away from zero and infinity, which implies that the convergence rate of the estimator $\tilde{\beta}_{ij}$ is of order $T^{-1/2}$, and the convergence rate of the estimator $\tilde{g}_i(u)$ is of order $T^{-2/5}$, which is as expected as we assumed the functions are twice differentiable.

To make the statement about the correct-fitting case neat, we assume that all m_i are of the same order and all m_{ij} are of the same order ($\max_{i,j} m_{ij}/\min_{i,j} m_{ij}$ and $\max_i m_i/\min_i m_i$ are bounded) as in (C5).

Theorem 2 *(Correct-fitting case). For any i, $i = 1, \cdots, m$, and $1 \le j \le p$, under the conditions (C1)-(C6), we have*

$$(mpT)^{1/2}(\mathbf{e}_{ij}^{\mathrm{T}}\mathbf{\Theta}_2\mathbf{e}_{ij})^{-1/2}\left(\hat{\beta}_{ij}-\beta_{ij}\right)\xrightarrow{D}N(0,\ 1)$$

and

$$(mT)^{2/5}(\mathbf{b}_i^{\mathrm{T}}(u)\mathbf{\Theta}_1\mathbf{b}_i(u))^{-1/2}\left(\hat{g}_i(u)-g_i(u)-r_i(u)\right)\xrightarrow{D}N(0,\ 1),$$

where $\mathbf{e}_{ij}$ and $\mathbf{b}_i(u)$ are unit vectors, $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$ are matrices with eigenvalues bounded and bounded away from zero, all these quantities are defined in the proof in the Appendix.

Theorem 2 shows the convergence rate of the estimator $\hat{\beta}_{ij}$ is of order $(mpT)^{-1/2}$, and the convergence rate of the estimator $\hat{g}_i(u)$ is of order $(mT)^{-2/5}$. This together with Theorem 1 show the estimators obtained by correct-fitting are of higher order of convergence rate than that obtained by over-fitting, therefore, more accurate. So, the homogeneity pursuit is imperative.

Because the asymptotic variance of either $\hat{\beta}_{ij}$ or $\hat{g}_i(u)$ has a very complicated form and it is not clear how to estimate it consistently, to construct a statistical inference based on the asymptotic normality established in Theorem 2 can be very challenging. In this paper we do not consider the inferences problem and leave it as an open question.

Let $\check{\boldsymbol{\beta}}_i$ and $\check{g}_i(\cdot)$ be the estimators of $\boldsymbol{\beta}_i$ and $g_i(\cdot)$ obtained under the assumption that all individuals share the same unknown parameters, which is the under-fitting case. We have

Theorem 3 *(Under-fitting case). Suppose the $\boldsymbol{\beta}_i$s are sufficiently separated in the sense that for $\bar{\boldsymbol{\beta}} := \sum_{i=1}^m \boldsymbol{\beta}_i/m$,*

$$\frac{1}{mp}\sum_{i=1}^{m}\|\boldsymbol{\beta}_i - \bar{\boldsymbol{\beta}}\|^2 \geq c$$

for some $c > 0$, then

$$\frac{1}{mp}\sum_{i=1}^{m}\|\check{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|^2 \geq c.$$

Similarly, if

$$\frac{1}{m}\sum_{i=1}^{m}\int |g_i(u) - \bar{g}(u)|^2\, du \geq c,$$

where $\bar{g}(u) = m^{-1}\sum_{i=1}^m g_i(u)$, then

$$\frac{1}{m}\sum_{i=1}^{m}\int |\check{g}_i(u) - g_i(u)|^2\, du \geq c.$$

Theorem 3 shows the estimators obtained by under-fitting are even not consistent, therefore, the worst.

4 Simulation studies

In this section, we are going to use simulated examples to assess the finite-sample performance of the proposed estimation.

Example 1. We generate a sample from model (1.2) with $p = 2$ and an even m,

$$g_i(u) = \begin{cases} \sin(\pi u/2) & \text{when } i = 1,\ 2,\ \cdots,\ m/2, \\ \cos(\pi u/2) & \text{when } i = m/2+1,\ \cdots,\ m, \end{cases}$$

and

$$\boldsymbol{\beta}_i = \begin{cases} (1,\ -1.5\sqrt{0.2},\ -0.5\sqrt{0.2})^{\mathrm{T}} & \text{when } i = 1,\ 3,\ \cdots,\ m-1, \\ (1,\ 0.5\sqrt{0.2},\ 1.5\sqrt{0.2})^{\mathrm{T}} & \text{when } i = 2,\ 4,\ \cdots,\ m, \end{cases}$$

where $\|\boldsymbol{\beta}_i\|^2 = 1.5$ for $i = 1,\ \cdots,\ m$. Let $\mathbf{X}_{it}$ and ϵ_{it}, $i = 1,\ \cdots,\ m$, $t = 1,\ \cdots,\ T$ be independently generated from $\frac{1}{\sqrt{1.5}}N(\mathbf{0}_3,\ \mathbf{I}_3)$ truncated by $[-1.343,\ 1.343]^3$ (the range of 5th to 95th quantiles for $N(0,\ 2/3)$) and $N(0,\ \sigma^2)$, respectively. Once $\mathbf{X}_{it}$ and ϵ_{it} are generated, y_{it} can be generated through (1.2).

We conduct this simulated example for various ms and Ts with $\sigma = 0.2$, and compare the proposed estimation with the potential competing methods based on the following performance metrics:

(1) *Estimation accuracy.* For an estimator $\hat{\boldsymbol{\beta}}_i$ of $\boldsymbol{\beta}_i$, we use the mean squared error (MSE), namely $\mathrm{MSE}(\hat{\boldsymbol{\beta}}_i) = E\left(\|\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i\|^2\right)$, to assess the estimation error of $\hat{\boldsymbol{\beta}}_i$. Analogously, for

an estimator $\hat{g}_i(\cdot)$ of $g_i(\cdot)$, its estimation accuracy can be evaluated based on the mean integrated squared error,

$$\text{MISE}(\hat{g}_i) = E\left\{\int \left(\hat{g}_i(u) - g_i(u)\right)^2 du\right\}.$$

To avoid the situation where the performance is dominated by the poor boundary behaviour, we let the integral domain to be non-boundary region, which is between the 1st and 99th quantiles of $\{\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i,\ t = 1,\ \cdots,\ T\}$.

(2) *Homogeneity structure identification consistency.* To evaluate the distance between the identified homogeneity structure and the true one, we use the normalized mutual information (NMI) (Ke et al.; 2015), which measures the similarity between two partitions. Suppose $\mathbb{C} = \{C_1,\ C_2,\ \cdots\}$ and $\mathbb{D} = \{D_1,\ D_2,\ \cdots\}$ are two partitions of $\{1,\ \cdots,\ n\}$, the NMI is defined as

$$\text{NMI}(\mathbb{C},\ \mathbb{D}) = \frac{I(\mathbb{C},\ \mathbb{D})}{[H(\mathbb{C}) + H(\mathbb{D})]/2},$$

where

$$I(\mathbb{C},\ \mathbb{D}) = \sum_{k,j} \big(|C_k \cap D_j|/n\big)\log\big(n|C_k \cap D_j|/|C_k||D_j|\big)$$

and

$$H(\mathbb{C}) = -\sum_{k} \big(|C_k|/n\big)\log\big(|C_k|/n\big).$$

The NMI takes values in $[0,\ 1]$ with larger values indicating higher level of similarity between two partitions. For an estimated partition $\hat{\mathbb{G}}_2 = \{\hat{G}_{2,1},\ \cdots,\ \hat{G}_{2,\hat{H}_2}\}$ of $\{(i,\ j) :\ 1,\ \cdots,\ m,\ j = 1,\ \cdots,\ p\}$, obtained in Stage 2 of the proposed estimation procedure in Section 2.1, we calculate $\text{NMI}(\hat{\mathbb{G}}_2,\ \mathbb{G}_2)$ to assess how close to the true homogeneity structure in β_{ij}s the estimated one is. Similarly, for an estimated partition $\hat{\mathbb{G}}_1$ of $\{i :\ 1,\ \cdots,\ m\}$, we use $\text{NMI}(\hat{\mathbb{G}}_1,\ \mathbb{G}_1)$ to evaluate how close the estimated homogeneity structure in $g_i(\cdot)$s is to the true one.

For each case, we apply either the single index model (1.2) with the initial estimation of the proposed estimation procedure in Section 2.1, which we call over-fitting (Over), the single index model (1.2) with the homogeneity structure (1.3) together with the proposed estimation procedure in Section 2.1, which we call correct-fitting, or the single index model (1.2) with all individuals sharing both the same index vector (namely, $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_m$) and the same link function (namely, $g_1(\cdot) = \cdots = g_m(\cdot)$, i.e. $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_m$), which we call under-fitting (Under), to the simulated data set. We also implemented two other kinds of under-fitting methods leading to similar empirical performance as Under, so we choose to put their results only in the Supplementary Material.

We develop three methods under the correct-fitting case. The first approach, named Correct-C, optimises (2.3) based on the estimated *componentwise* homogeneity structure in β_{ij}s and θ_{ij}s, obtained in Stage 2 of the estimation procedure in Section 2.1 with the tuning parameters selected by the 10-fold cross-validation procedure described in Section 2.3. The second approach,

Correct-V, is the same as the first approach but optimises (2.3) based on the estimated componentwise homogeneity structure in β_{ij}s and *vectorwise* homogeneity structure in $\boldsymbol{\theta}_i$s which is obtained based on the estimators of θ_{ij}s obtained under the estimated componentwise homogeneity structure in θ_{ij}s. The third approach, which we call Correct-NMI, is the same as the second approach but with the tuning parameters selected to be those maximising NMI($\hat{\mathbb{G}}_2,\ \mathbb{G}_2$) and NMI($\hat{\mathbb{G}}_1,\ \mathbb{G}_1$). In practice without knowing the true homogeneity structure, one cannot implement Correct-NMI. We compare over-fitting, correct-fittings and under-fittings to the oracle case where the true homogeneity structure is used. The computational algorithms for the under-fittings and oracle estimators are the same as that for the correct-fittings, but use either the identified or pre-specified homogeneity structure.

We also adopt variants of the k-means-based approach (Vogt and Linton; 2017) by first clustering link functions and index coefficients based on their initial estimates and then averaging the initial estimates within each cluster to produce the final estimates. Specifically, after Stage 1 of the estimation procedure in Section 2.1 where we can obtain $\tilde{g}_i(\cdot) = \mathbf{B}(\cdot)^{\mathrm{T}}\tilde{\boldsymbol{\theta}}_i$ and $\tilde{\beta}_{ij}$ for $i = 1, \cdots, m, j = 1, \cdots, p$, we implement the k-means algorithm to cluster either only $\tilde{g}_i(\cdot)$s based on the L^2 distance, which we call Km-F or both $\tilde{\beta}_{ij}$s and $\tilde{g}_i(\cdot)$s, which we call Km-I-F. The corresponding tuning parameters, i.e. numbers of clusters in $\tilde{\beta}_{ij}$s and $\tilde{g}_i(\cdot)$s, are still selected by the cross-validation approach. We investigate the sample performance of all ten approaches in the simulation study.

We report the results for estimation errors and NMIs for $\boldsymbol{\beta}_i$s and $g_i(\cdot)$s averaged over 100 replicates in Tables 1 and 2, respectively. In terms of the estimation error, the overall estimation accuracy is improved as m and T increase and correct-fittings and Km-I-F perform very well as reflected in their lower values of MSEs and MISEs. Among three correct-fitting methods, Correct-NMI provides the best performance even producing very comparable MSEs and MISEs with the oracle estimator and Correct-C is outperformed by Correct-V in terms of MISEs in all settings. Intuitively, this is because, unlike Correct-C optimising (2.3) based on the detected homogeneity structure in β_{ij}s and θ_{ij}s, Correct-V separates the final estimation step from the cross-validation procedure, which is used to identify the homogeneity structure in β_{ij}s and $\boldsymbol{\theta}_i$s. Analogously, Correct-NMI solves a separate optimisation after detecting the homogeneity structure based on the largest NMIs. Moreover, in comparison with k-means-based methods, correct-fittings are substantially superior with lower MSE values in most settings, but are outperformed in terms of MISEs in many settings due to relatively smaller NMI values especially when T is not large enough. As long as T becomes sufficiently large, Correct-V and Correct-NMI provide prominently improved estimates over Km-F and Km-I-F with respect to lower values of MISEs. It is also worth noting that the over-fitting and under-fitting methods, which either ignores or mistakenly specify the homogeneity structure, provide much worse results, highlighting the importance of incorporating the appropriate homogeneity structure.

In terms of the homogeneity structure selection consistency, we observe that three correct-

Table 1: **The Average of MSE($\hat{\boldsymbol{\beta}}_i$), $i = 1, \cdots, m$, and Average NMIs for $\hat{\mathbb{G}}_2$**

All entries for MSEs are 10^4 times their actual values

	T	200			400			800		
	m	30	60	90	30	60	90	30	60	90
MSE	Oracle	0.539	0.389	0.204	0.317	0.190	0.121	0.191	0.121	0.097
	Correct-C	0.895	0.633	0.341	0.434	0.212	0.143	0.166	0.088	0.091
	Correct-V	0.547	0.401	0.214	0.320	0.189	0.122	0.191	0.122	0.096
	Correct-NMI	0.539	0.386	0.205	0.318	0.189	0.121	0.191	0.121	0.097
	Over	10.461	10.641	10.520	5.361	5.214	5.188	2.652	2.698	2.632
	Under	4015.5	4006.5	4004.9	4005.7	4003.5	4002.2	4003.5	4001.8	4001.1
	Km-F	10.474	10.708	10.555	5.340	5.222	5.199	2.655	2.712	2.629
	Km-I-F	0.827	0.440	0.233	0.526	0.233	0.128	0.300	0.137	0.085
NMI	Oracle	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correct-C	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correct-V	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correct-NMI	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Over	0.506	0.449	0.421	0.506	0.449	0.421	0.506	0.449	0.421
	Under	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.667
	Km-F	0.506	0.449	0.421	0.506	0.449	0.421	0.506	0.449	0.421
	Km-I-F	0.997	1.000	1.000	0.996	0.999	1.000	0.993	0.999	1.000

fitting methods produce perfect identifications of the homogeneity structure in β_{ij}s and k-means based methods provide the most accurate recovery of $\mathbb{G}_1$ when $T = 200$ or 400. For $T = 800$, both Correct-NMI and k-means methods can almost perfectly identify the homogeneity structure in $g_i(\cdot)$s. Moreover, the performance of Correct-C and Correct-V deteriorates when m increases, this is intuitively due to the increased m values and the cross-validation procedure, which tends to choose a larger number of change points as m increases, resulting in smaller NMI values for $\hat{\mathbb{G}}_1$.

Table 2: **The Average of MISE**$(\hat{g}_i)$**,** $i = 1, \cdots, m$**, and Average NMIs for** $\hat{\mathbb{G}}_1$

All entries for MISEs are 10^2 *times their actual values*

	T	200			400			800		
	m	30	60	90	30	60	90	30	60	90
MISE	Oracle	0.278	0.260	0.248	0.257	0.248	0.246	0.246	0.237	0.238
	Correct-C	1.137	1.108	1.030	0.652	0.542	0.521	0.364	0.307	0.312
	Correct-V	0.517	0.569	0.635	0.290	0.281	0.300	0.251	0.241	0.243
	Correct-NMI	0.313	0.295	0.278	0.262	0.253	0.251	0.246	0.237	0.237
	Over	0.813	0.805	0.801	0.543	0.544	0.543	0.404	0.396	0.397
	Under	86.670	86.545	86.461	87.955	87.925	87.946	88.700	88.697	88.661
	Km-F	0.281	0.253	0.236	0.264	0.248	0.246	0.261	0.246	0.244
	Km-I-F	0.281	0.251	0.236	0.263	0.247	0.245	0.260	0.246	0.244
NMI	Oracle	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Correct-C	0.537	0.398	0.314	0.844	0.806	0.669	0.961	0.951	0.902
	Correct-V	0.537	0.398	0.314	0.844	0.806	0.669	0.961	0.951	0.902
	Correct-NMI	0.877	0.836	0.835	0.970	0.963	0.950	0.994	0.998	0.997
	Over	0.339	0.290	0.267	0.339	0.290	0.267	0.339	0.290	0.267
	Under	0	0	0	0	0	0	0	0	0
	Km-F	0.990	0.993	1.000	0.988	0.996	0.998	0.987	1.000	0.998
	Km-I-F	0.990	0.997	1.000	0.991	1.000	1.000	0.993	1.000	1.000

In Example 1 where $g_i(\cdot)$s are generated via sine and cosine functions, it is apparent that those vectorwise-homogeneity-structure-based methods, including Correct-V, Correct-NMI, Km-F and Km-I-F, provide more accurate estimates of $g_i(\cdot)$s than Correct-C. As long as the link functions are generated in a componentwise fashion, we will use Example 2 to demonstrate the superiority of Correct-C to its competitors. See Section S.4 of the Supplementary Material for full details.

5 Real data analysis

We will illustrate the proposed method with two real data examples in this section.

5.1 Industrial Portfolio's return

We first study the data set about $m = 49$ Industrial Portfolios' daily simple return from 1/8/2015 to 31/12/2015. This data set can be freely downloaded from Kenneth French's website `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html` This data set has been analysed in quite a few literature. For example, Guo et al. (2017) used this data set to demonstrate the performance of a newly developed dynamic portfolio allocation. In this paper, we are going to explore the homogeneity structure in this data set by our proposed method.

Let y_{it} be the daily simple return of the ith portfolio at the tth day, $i = 1, \cdots, m$, $t = 1, \cdots, T$, and $\mathbf{X}_{it} = (X_{t1}, X_{t2}, X_{t3})^{\mathrm{T}}$ be the observation of the Fama-French three factors, where X_{t1}, X_{t2}, X_{t3} respectively represent the market (Rm-Rf), size (SMB) and value (HML) factors at the tth day.

We apply the single index model (1.2) with the unknown homogeneity structure (1.3) to fit the data set. From interpretation point of view, the homogeneity structure in the unknown link functions, $g_i(\cdot)$s, where each $g_i(\cdot)$ is treated as a single undivided unit would make much more sense than the homogeneity structure in the coefficients of the B-Spline decompositions of $g_i(\cdot)$s. Therefore, we use the Correct-V, described in Section 4, to identify the homogeneity structure in β_{ij}s or $g_i(\cdot)$s, and estimate the unknown parameters and unknown functions.

In the implementation of the Correct-V, we apply the method in Section 2.1 with the tuning parameters selected by the cross-validation for time dependent data as described in Section 2.3. Specifically, we define the cross-validated mean squared error

$$\mathrm{CV} = \frac{1}{mL} \sum_{i=1}^{m} \sum_{t=T-L+1}^{T} (y_{it} - \hat{y}_{it})^2 \tag{5.1}$$

where $L = 30$. Note that we here do not apply the one-standard-rule when performing the cross validation to select the tuning parameters for identifying the homogeneity structure, since we have already selected a small enough model with 11 and 2 detected groups in index coefficients and link functions, respectively. Table 3 provides the identified clustering results for β_{i2}, β_{i3}, $g_i(\cdot)$, $i = 1, \cdots, 49$. Table 4 presents the estimated index coefficients and Figure 1 plots the estimated link functions. We observe a few apparent patterns. Firstly, the estimated link functions are very linear indicating the linear relationship between portfolio returns and Fama-French three factors, which has been verified by broad empirical studies. Secondly, many portfolios belonging to similar industrials were grouped into the same cluster for the estimated index coefficients, e.g. Hardw, Softw and Agric, Food, Soda were clustered into Groups 8 and 4 in terms of the estimated coefficients for factors SMB and HML respectively.

5.2 UK climate data

Our second data set, which is available from the UK Met Office website

Table 3: **Grouping Results for The Index Coefficients for SMB, HML and Link Functions of 49 Industrial Portfolios**

	Agric	Food	Soda	Beer	Smoke	Toys	Fun	Books	Hshld	Clths
SMB	6	5	6	4	2	8	9	7	6	6
HML	4	4	4	3	4	5	3	4	5	6
Function	i	ii	i	i	ii	ii	ii	ii	ii	i
	Hth	MedEq	Drugs	Chems	Rubbr	Txtls	BldMt	Cnstr	Steel	FabPr
SMB	9	8	9	7	7	8	8	8	9	10
HML	2	2	1	6	4	4	6	6	9	7
Function	ii	ii	ii	ii	ii	ii	ii	ii	ii	i
	Mach	ElcEq	Autos	Aero	Ships	Guns	Gold	Mines	Coal	Oil
SMB	8	9	7	6	8	6	10	9	11	9
HML	8	6	6	6	7	4	10	7	11	9
Function	ii	ii	ii	ii	ii	ii	i	ii	i	ii
	Util	Telcm	PerSv	BusSv	Hardw	Softw	Chips	LabEq	Paper	Boxes
SMB	4	8	8	7	8	8	7	7	6	6
HML	5	4	5	4	5	3	5	5	5	6
Function	ii	ii	ii	ii	ii	ii	ii	ii	ii	ii
	Trans	Whlsl	Rtail	Meals	Banks	Insur	RIEst	Fin	Other	
SMB	7	7	7	7	7	6	7	6	7	
HML	6	6	4	3	6	5	5	5	6	
Function	ii	ii	ii	ii	ii	ii	ii	ii	i	

Table 4: **Estimated Index Coefficients.**

$\hat{\beta}_{(1)}$	$\hat{\beta}_{(2)}$	$\hat{\beta}_{(3)}$	$\hat{\beta}_{(4)}$	$\hat{\beta}_{(5)}$	$\hat{\beta}_{(6)}$	$\hat{\beta}_{(7)}$	$\hat{\beta}_{(8)}$	$\hat{\beta}_{(9)}$	$\hat{\beta}_{(10)}$	$\hat{\beta}_{(11)}$
-0.594	-0.203	-0.087	0.056	0.134	0.259	0.393	0.462	0.599	0.657	1.456

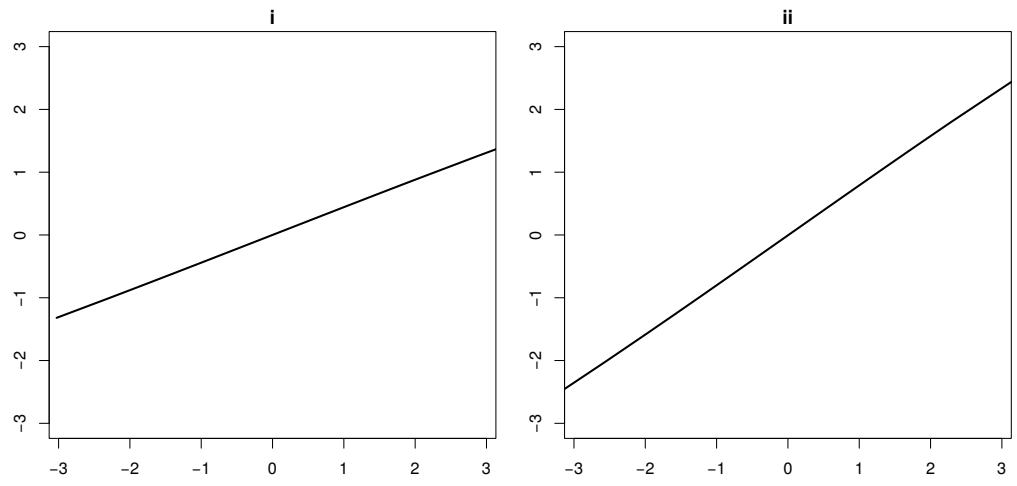

Figure 1: Plots of estimated link functions.

`http://www.metoffice.gov.uk/public/weather/climate-historic`,

contains monthly data of the mean daily maximum temperature (TMAX), days of air frost (AF), total rainfall (RAIN) and total sunshine duration (SUN) collected from 37 stations across the UK. We first remove the missing values and thus select data during the period of January 1993 to December 2009 from 16 locations. We then eliminate the seasonality, trend effects and standardise the data. Let y_{it} and $\mathbf{X}_{it} = (X_{it1},\ X_{it2},\ X_{it3})^{\mathrm{T}}$ be the observations for TMAX, AF, RAIN and SUN, respectively, from the ith station at the tth month, $i = 1, \cdots, 16$, $t = 1, \cdots, 204$.

Like the analysis of the Industrial Portfolio's return data set, we apply the single index model (1.2) with unknown homogeneity structure (1.3) together with the proposed estimation procedure, Correct-V, to the data set. Table 5 provides the clustering results for the index coefficients, where 5 groups were selected with estimated coefficients $\hat{\beta}_{(1)} = -0.8$, $\hat{\beta}_{(2)} = -0.634$, $\hat{\beta}_{(3)} = -0.359$, $\hat{\beta}_{(4)} = 0.012$ and $\hat{\beta}_{(5)} = 0.173$. Moreover, only one group was selected while clustering the link functions. See Figure 2 for the plot of this estimated link function.

The estimated function has a reverse S-shape and exhibits a decreasing trend with a faster decay rate in the middle. It is easy to observe, from Table 5, that the estimated coefficients for rainfall and sunshine duration belong to Groups 4–5 and Groups 1–3, respectively. To summarize, days of air frost and total sunshine duration have apparently negative and positive impacts on the monthly maximum temperature, respectively. Nevertheless, the negative impact of rainfall is quite weak. It is also very interesting to see that Sheffield, Shawbury, Ross-On-Wye, Oxford, Heathrow, Eskdalemuir, Bradford share exactly the same model, which implies the impact of rainfall or total sunshine duration on monthly maximum temperature has exactly the same pattern in these seven areas. The same finding also appears in three areas of Lerwick, Eastbourne and Camborne, the two areas of Waddington and Cambridge, and the two areas of

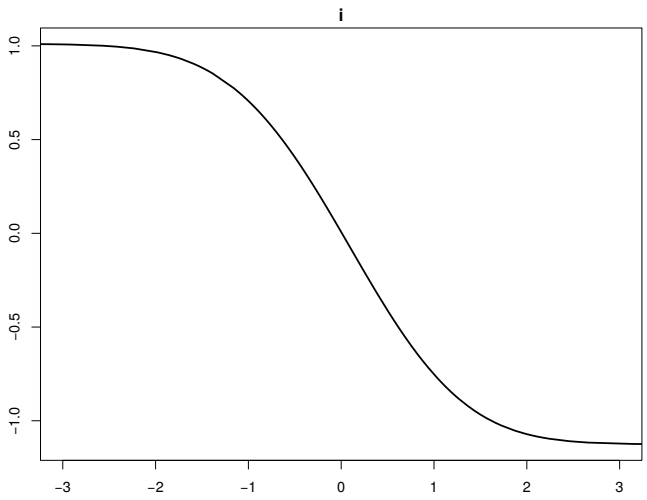

Figure 2: Plot of the estimated link function.

Paisley and Leuchars.

Table 5: **Grouping Results for the Index Coefficients for RAIN, SUN at 16 Locations**

	Waddington	Sheffield	Shawbury	Ross-On-Wye	Paisley	Oxford	Leuchars	Lerwick
RAIN	4	4	4	4	5	4	5	4
SUN	1	2	2	2	3	2	3	3
	Hurn	Heathrow	Eskdalemuir	Eastbourne	Cambridge	Camborne	Bradford	Armagh
RAIN	5	4	4	4	4	4	4	5
SUN	1	2	2	3	1	3	2	2

Acknowledgements

The authors sincerely thank the Editor Professor Christian Hansen, the Associate Editor and two anonymous reviewers for their insightful comments that significantly improve the paper. The research of Heng Lian is supported by Hong Kong RGC general research fund 11301718, and by Project 11871411 from NSFC and the Shenzhen Research Institute, City University of Hong Kong. This research is also supported by National Natural Science Foundation of China (Grant Number 71833004)

Appendix

The Appendix includes additional notations and brief technical proofs supporting Section 3 in Sections A.1 and A.2–A.5, respectively.

A.1 Additional notations

Let $\epsilon_t = (\epsilon_{1t}, \cdots, \epsilon_{mt})^{\mathrm{T}}$, $y_t = (y_{1t}, \cdots, y_{mt})^{\mathrm{T}}$, $\mathbf{X}_t = (\mathbf{X}_{1t}^{\mathrm{T}}, \cdots, \mathbf{X}_{mt}^{\mathrm{T}})^{\mathrm{T}}$. Due to assumption (C3), there exists $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_{0m}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\theta}_{0i} = (\theta_{0i1}, \ldots, \theta_{0iK})^{\mathrm{T}}$ such that $\sup_x |g_{0i}(x) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(x)| \leq CK^{-2}$. Here and below we use C to denote a generic positive constant whose value can change even on the same line. We use $\|.\|_{op}$ to denote the operator norm of a matrix (the operator norm is the same as the largest singular value) and use $\|.\|$ to denote the Frobenius norm of a matrix. We use $\|.\|_{L^2}$ to denote the L^2 norm of functions and $\|.\|_\infty$ is the sup-norm for vectors (maximum absolute value of the components).

Assume the true partition of components of $\boldsymbol{\theta}_0$ and $\overline{\boldsymbol{\beta}}_0$ is given by $\cup_{h=1}^{H_1} G_{1,h} = \{1, \ldots, mK\}$ and $\cup_{h=1}^{H_2} G_{2,h} = \{1, \ldots, mp\}$, respectively. The unique values of the components of $\boldsymbol{\theta}_0$ and $\overline{\boldsymbol{\beta}}$ are denoted by $\boldsymbol{\xi}_0 = (\xi_{01}, \ldots, \xi_{0H_1})^{\mathrm{T}} \in R^{H_1}$ and $\boldsymbol{\eta}_0 = (\eta_{01}, \ldots, \eta_{0H_2})^{\mathrm{T}} \in R^{H_2}$, respectively. Let $\mathbf{J}_i^{G_1}$ be the $K \times H_1$ binary matrix whose (k, h) entry is 1 if $\theta_{0ik} = \xi_h$ and 0 otherwise. We have $\boldsymbol{\theta}_{0i} = \mathbf{J}_i^{G_1}\boldsymbol{\xi}_0$. Similarly, we define $\mathbf{J}_i^{G_2}$ such that $\boldsymbol{\beta}_{0i} = \mathbf{J}_i^{G_2}\boldsymbol{\eta}_0$. The sizes of $G_{1,h}$ and $G_{2,h}$ are denoted by $|G_{1,h}|$ and $|G_{2,h}|$. Finally, let $\mathbf{D}^{G_1}$ and $\mathbf{D}^{G_2}$ be the diagonal matrix with entries $\sqrt{|G_{1,h}|}$ and $\sqrt{|G_{2,h}|}$, respectively.

A.2 Proof summary

We first define the oracle estimator as the minimizer $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ of

$$\min_{\boldsymbol{\theta}, \overline{\boldsymbol{\beta}}} \sum_{i=1}^{m} \sum_{t=1}^{T} (y_{it} - \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i)\boldsymbol{\theta}_i)^2,$$

where $\boldsymbol{\beta}_i = (1, \overline{\boldsymbol{\beta}}_i^{\mathrm{T}})^{\mathrm{T}} = (1, \beta_{i1}, \ldots, \beta_{ip})^{\mathrm{T}}$ and $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{iK})^{\mathrm{T}}$ with the constraint that components of $\overline{\boldsymbol{\beta}} = (\overline{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \overline{\boldsymbol{\beta}}_m^{\mathrm{T}})^{\mathrm{T}}$ in the same partition take the same value and components of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_m^{\mathrm{T}})^{\mathrm{T}}$ in the same partition take the same value. Here we assume the partition is the true partition, thus the name "oracle".

Below we first show that the oracle estimator satisfies the asymptotic normality properties stated in Theorem 2 (we also obtained convergence rate and asymptotic normality for the entire vector $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, see for example (A.9) and (A.10). Also, noting that all arguments carry over when the partition used in the oracle estimator is finer than the true partition, Theorem 1 follows directly as a special case that each component of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ forms its own group in the partition. Then we show that the change points can be consistently estimated, and thus the estimator we obtain in stage 3 will be exactly the same as the oracle estimator using the true partition, with probability approaching one, and Theorem 2 is proved. The rest of the appendix contains a sketch of the proofs outlined above while more details are relegated to the supplementary material, as well as several lemmas.

A.3 Proof of asymptotic property for the oracle estimator

In this part we consider the asymptotic property of the oracle estimator, denoted by $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ in this section, which assumed knowledge of the true partitions. For clarity of presentation, the proof is split into several steps.

STEP 1. Prove the convergence rate $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| + \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\sqrt{(H_1 + H_2)/T} + \sqrt{m}K^{-2})$.

In this section, when we use $\boldsymbol{\theta}$, we always assume $\boldsymbol{\theta}_i = \mathbf{J}_i^{G_1}\boldsymbol{\xi}$ for some $\boldsymbol{\xi} \in R^{H_1}$ (that is, components of $\boldsymbol{\theta}$ are partitioned in the same way as is the true $\boldsymbol{\theta}_0$). It is easy to see that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = \|\mathbf{D}^{G_1}(\boldsymbol{\xi} - \boldsymbol{\xi}_0)\|$. Similarly, we always assume $\boldsymbol{\beta}_i = \mathbf{J}_i^{G_2}\boldsymbol{\eta}$ for some $\boldsymbol{\eta} \in R^{H_2}$ and $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = \|\mathbf{D}^{G_2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0)\|$.

Define $r_T = \sqrt{(H_1 + H_2)/T} + \sqrt{m}K^{-2}$. We only need to show that

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2+\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2=Lr_T^2} \sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i))^2 - \sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2 > 0$$

with probability approaching one, if L is large enough.

We have

$$\begin{aligned}
&\sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i))^2 - \sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2 \\
= &\sum_{i,t}(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2 - 2(\epsilon_{it} - r_{it})(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})),
\end{aligned}$$

where $r_{it} = \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})$ with $|r_{it}| \le CK^{-2}$.

Furthermore, some algebra shows

$$\begin{aligned}
&\sum_{i,t}(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2 \\
= \; & T((\boldsymbol{\xi}^{\mathrm{T}} - \boldsymbol{\xi}_0^{\mathrm{T}})\mathbf{D}^{G_1}, (\boldsymbol{\eta}^{\mathrm{T}} - \boldsymbol{\eta}_0^{\mathrm{T}})\mathbf{D}^{G_2}) \\
& \begin{pmatrix} (\mathbf{D}^{G_1})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{D}^{G_2})^{-1} \end{pmatrix} \begin{pmatrix} (\mathbf{J}_1^{G_1})^{\mathrm{T}} & \mathbf{0} & \cdots & (\mathbf{J}_m^{G_1})^{\mathrm{T}} & \mathbf{0} \\ \mathbf{0} & (\mathbf{J}_1^{G_2})^{\mathrm{T}} & \cdots & \mathbf{0} & (\mathbf{J}_m^{G_2})^{\mathrm{T}} \end{pmatrix} \\
& \begin{pmatrix} \widetilde{\mathbf{A}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{A}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widetilde{\mathbf{A}}_m \end{pmatrix} \cdot \begin{pmatrix} \mathbf{J}_1^{G_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_1^{G_2} \\ \vdots & \vdots \\ \mathbf{J}_m^{G_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_m^{G_2} \end{pmatrix} \begin{pmatrix} (\mathbf{D}^{G_1})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{D}^{G_2})^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{D}^{G_1}(\boldsymbol{\xi} - \boldsymbol{\xi}_0) \\ \mathbf{D}^{G_2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0) \end{pmatrix},
\end{aligned}$$

where

$$\widetilde{\mathbf{A}}_i = \frac{1}{T}\sum_{t=1}^{T}\left[\begin{pmatrix} \mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) \\ \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i^*)\overline{\mathbf{X}}_{it} \end{pmatrix} \begin{pmatrix} \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) & \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i^*)\overline{\mathbf{X}}_{it}^{\mathrm{T}} \end{pmatrix}\right], 1 \le i \le m,$$

$\mathbf{B}'(.) = (B_1'(.), \ldots, B_K'(.))^{\mathrm{T}}$ are the first derivatives of the basis functions and $\boldsymbol{\beta}_i^*$ lies between $\boldsymbol{\beta}_{0i}$ and $\boldsymbol{\beta}_i$.

With the help of Lemma 3 in the SUPPLEMENTARY MATERIAL provided in a separated file, and noting that

$$\mathbf{O} := \begin{pmatrix} \mathbf{J}_1^{G_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_1^{G_2} \\ \vdots & \vdots \\ \mathbf{J}_m^{G_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_m^{G_2} \end{pmatrix} \begin{pmatrix} (\mathbf{D}^{G_1})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{D}^{G_2})^{-1} \end{pmatrix} \tag{A.1}$$

is an orthonormal matrix (that is, $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{I}$), we can get

$$\sum_{i,t}(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i)-\boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2 \asymp T(\|\mathbf{D}^{G_1}(\boldsymbol{\xi}-\boldsymbol{\xi}_0)\|^2+\|\mathbf{D}^{G_2}(\boldsymbol{\eta}-\boldsymbol{\eta}_0)\|^2) = T(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2). \tag{A.2}$$

Now consider the term $(\epsilon_{it} - r_{it})(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))$. We can show

$$\begin{aligned} & \sum_{i,t} \epsilon_{it}(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})) \\ \leq \quad & \sqrt{\|\mathbf{D}^{G_1}(\boldsymbol{\xi} - \boldsymbol{\xi}_0)\|^2 + \|\mathbf{D}^{G_2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0)\|^2} \left\| \sum_t \mathbf{O}^{\mathrm{T}} \begin{pmatrix} \mathbf{B}(\mathbf{X}_{1t}^{\mathrm{T}}\boldsymbol{\beta}_1)\epsilon_{1t} \\ \boldsymbol{\theta}_{01}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{1t}^{\mathrm{T}}\boldsymbol{\beta}_1^*)\overline{\mathbf{X}}_{1t}\epsilon_{1t} \\ \vdots \\ \mathbf{B}(\mathbf{X}_{mt}^{\mathrm{T}}\boldsymbol{\beta}_m)\epsilon_{mt} \\ \boldsymbol{\theta}_{0m}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{mt}^{\mathrm{T}}\boldsymbol{\beta}_m^*)\overline{\mathbf{X}}_{mt}\epsilon_{mt} \end{pmatrix} \right\|, \end{aligned}$$

where $\mathbf{O}$ is as defined in (A.1), and further calculations reveal that

$$\begin{aligned} & E \left\| \sum_t \mathbf{O}^{\mathrm{T}} \begin{pmatrix} \mathbf{B}(\mathbf{X}_{1t}^{\mathrm{T}}\boldsymbol{\beta}_1)\epsilon_{1t} \\ \boldsymbol{\theta}_{01}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{1t}^{\mathrm{T}}\boldsymbol{\beta}_1^*)\overline{\mathbf{X}}_{1t}\epsilon_{1t} \\ \vdots \\ \mathbf{B}(\mathbf{X}_{mt}^{\mathrm{T}}\boldsymbol{\beta}_m)\epsilon_{mt} \\ \boldsymbol{\theta}_{0m}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{mt}^{\mathrm{T}}\boldsymbol{\beta}_m^*)\overline{\mathbf{X}}_{mt}\epsilon_{mt} \end{pmatrix} \right\|^2 \\ \leq \quad & \mathrm{tr}(\mathbf{O}\mathbf{O}^{\mathrm{T}}) \cdot \left\| \sum_{1\leq t,t'\leq T} \begin{bmatrix} \mathbf{A}_{1,|t-t'|} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{2,|t-t'|} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_{m,|t-t'|} \end{bmatrix} \right\|_{op}, \end{aligned} \tag{A.3}$$

where

$$\mathbf{A}_{i,|t-t'|} = E\left[\begin{pmatrix} \mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) \\ \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i^*)\overline{\mathbf{X}}_{it} \end{pmatrix} \begin{pmatrix} \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it'}^{\mathrm{T}}\boldsymbol{\beta}_i) & \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{it'}^{\mathrm{T}}\boldsymbol{\beta}_i^*)\overline{\mathbf{X}}_{it'}^{\mathrm{T}} \end{pmatrix} \epsilon_{it}\epsilon_{it'} \right].$$

By Lemma 4 in the SUPPLEMENTARY MATERIAL provided in a separated file, and that

$\text{tr}(\mathbf{O}\mathbf{O}^{\mathrm{T}}) = H_1 + H_2$ (note $\mathbf{O}^{\mathrm{T}}\mathbf{O} = \mathbf{I}_{H_1+H_2}$), we have

$$\begin{aligned} & \sum_{i,t} \epsilon_{it}(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})) \\ = \; & O_p(\sqrt{(\|\boldsymbol{\theta}-\boldsymbol{\theta}\|^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2)(H_1+H_2)T}). \end{aligned} \tag{A.4}$$

Finally, using Cauchy-Schwarz inequality

$$\begin{aligned} & \sum_{i,t} r_{it}(\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})) \\ = \; & C\sqrt{mT}K^{-2} \cdot O_p(\sqrt{T(\|\boldsymbol{\theta}-\boldsymbol{\theta}\|^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2)}) \end{aligned} \tag{A.5}$$

Combining (A.2)–(A.5),

$$\sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i))^2 - \sum_{i=1}^{m}\sum_{t=1}^{T}(y_{it} - \boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2 > 0$$

with probability approaching one, if $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 + \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 = Lr_T^2$ with L sufficiently large. Thus there is a local minimizer $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ with $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| + \|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\| = O_p(r_T)$.

STEP 2. Proof of convergence rate of $\widehat{\boldsymbol{\beta}}$ and its asymptotic normality.

Let $\boldsymbol{\Pi}_i$ be $T \times K$ matrices, $i = 1, \ldots, m$, with rows $\boldsymbol{\Pi}_{it}^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})$. Define $\mathbf{V}_{it} = g_{0i}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\overline{\mathbf{X}}_{it}$, $\mathbf{P}_i = \boldsymbol{\Pi}_i(\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\Pi}_i)^{-1}\boldsymbol{\Pi}_i^{\mathrm{T}}$ with rows $\mathbf{P}_{it}^{\mathrm{T}} = \boldsymbol{\Pi}_{it}^{\mathrm{T}}(\boldsymbol{\Pi}_i^{\mathrm{T}}\boldsymbol{\Pi}_i)^{-1}\boldsymbol{\Pi}_i^{\mathrm{T}}$. We write, for any $(\boldsymbol{\theta}, \boldsymbol{\beta})$ with $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 \le Cr_T^2$ and $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 \le CH_2/T$,

$$\begin{aligned} & \sum_{i,t}(y_{it} - \boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i))^2 \\ = \; & \sum_{i,t}(\epsilon_{it} - \boldsymbol{\Pi}_{it}^{\mathrm{T}}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}) - \mathbf{V}_{it}^{\mathrm{T}}(\overline{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}}_{0i}) - R_{it})^2, \end{aligned}$$

where

$$\begin{aligned} & R_{it} \\ = \; & \left\{\boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\right\} + \left\{(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i})^{\mathrm{T}}(\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))\right\} \\ & + \left\{\boldsymbol{\theta}_{0i}^{\mathrm{T}}(\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i) - \mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - \mathbf{B}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\overline{\mathbf{X}}_{it}^{\mathrm{T}}(\overline{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}}_{0i})\right\} \\ & + \left\{(\boldsymbol{\theta}_{0i}^{\mathrm{T}}\mathbf{B}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g_{0i}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))\overline{\mathbf{X}}_{it}^{\mathrm{T}}(\overline{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}}_{0i})\right\} \\ = \; & R_{it1} + R_{it2}(\boldsymbol{\theta}_i, \boldsymbol{\beta}_i), \end{aligned}$$

with $R_{it1} = \boldsymbol{\theta}_0^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})$ and $R_{it2}(\boldsymbol{\theta}_i, \boldsymbol{\beta}_i)$ (or R_{it2} for short) contains the remaining terms. Using $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 + \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|^2 \le Cr_T^2$, we can show

$$\sum_{i,t} R_{it2}^2 = O_p\left(Tr_T^4K^3 + Tr_T^2K^{-2}\right). \tag{A.6}$$

We then orthogonalize the parametric part with respect to the nonparametric part by writing

$$\sum_{i,t}(\epsilon_{it} - \boldsymbol{\Pi}_{it}^{\mathrm{T}}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}) - \mathbf{V}_{it}^{\mathrm{T}}(\overline{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}}_{0i}) - R_{it})^2$$

$$= \sum_{i,t}(\epsilon_{it} - \mathbf{\Pi}_{it}^{\mathrm{T}}(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_{0i}) - (\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})^{\mathrm{T}}(\overline{\boldsymbol{\beta}}_i - \overline{\boldsymbol{\beta}}_{0i}) - R_{it1} - R_{it2}(\mathcal{M}_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i))^2,$$

where $\boldsymbol{\alpha}_i = \boldsymbol{\theta}_i + (\mathbf{\Pi}_i^{\mathrm{T}}\mathbf{\Pi}_i)^{-1}\mathbf{\Pi}_i^{\mathrm{T}}\mathbf{V}_i\overline{\boldsymbol{\beta}}_i$, $\boldsymbol{\alpha}_{0i} = \boldsymbol{\theta}_{0i} + (\mathbf{\Pi}_i^{\mathrm{T}}\mathbf{\Pi}_i)^{-1}\mathbf{\Pi}_i^{\mathrm{T}}\mathbf{V}_i\overline{\boldsymbol{\beta}}_{0i}$, $\mathbf{V}_i = (\mathbf{V}_{it}, \ldots, \mathbf{V}_{iT})^{\mathrm{T}}$, and $\mathcal{M}_i$ is the one-to-one mapping that maps $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ to $(\boldsymbol{\theta}_i, \boldsymbol{\beta}_i)$. Below we write $R_{it2}(\mathcal{M}_i(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i))$ as R_{it2}, $R_{it2}(\mathcal{M}_i(\widehat{\boldsymbol{\alpha}}_i, \widehat{\boldsymbol{\beta}}_i))$ as $\widehat{R}_{it2}$, and note $R_{it2}(\mathcal{M}_i(\widehat{\boldsymbol{\alpha}}_i, \boldsymbol{\beta}_0)) = 0$. Then,

$$\begin{aligned}
0 \;\geq\; & \sum_{i,t}(\epsilon_{it} - \mathbf{\Pi}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_i - (\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})^{\mathrm{T}}(\widehat{\overline{\boldsymbol{\beta}}}_i - \overline{\boldsymbol{\beta}}_{0i}) - R_{it1} - R_{it2})^2 \\
& - \sum_{i,t}(\epsilon_{it} - \mathbf{\Pi}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_i - R_{it1})^2 \\
= \;& \sum_{i,t}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^{\mathrm{T}}(\mathbf{J}_i^{G_2})^{\mathrm{T}}(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})(\mathbf{V}_{it}^{\mathrm{T}} - \mathbf{P}_{it}^{\mathrm{T}}\mathbf{V}_i)\mathbf{J}_i^{G_2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \\
& - 2\sum_{i,t}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^{\mathrm{T}}(\mathbf{J}_i^{G_2})^{\mathrm{T}}(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})\epsilon_{it} \\
& - 2\sum_{i,t}\widehat{R}_{it2}\epsilon_{it} \\
& + \sum_{i,t}\widehat{R}_{it2}^2 + 2\sum_{i,t}\left((\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^{\mathrm{T}}(\mathbf{J}_i^{G_2})^{\mathrm{T}}(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it}) + \widehat{R}_{it2}\right)(\mathbf{\Pi}_{it}^{\mathrm{T}}\widehat{\boldsymbol{\alpha}}_i + R_{it1}) \\
& + 2\sum_{i,t}\widehat{R}_{it2} \cdot (\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^{\mathrm{T}}(\mathbf{J}_i^{G_2})^{\mathrm{T}}(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it}).
\end{aligned} \tag{A.7}$$

The first term above is

$$\begin{aligned}
& \sum_{i,t}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)(\mathbf{J}_i^{G_2})^{\mathrm{T}}(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})(\mathbf{V}_{it}^{\mathrm{T}} - \mathbf{P}_{it}^{\mathrm{T}}\mathbf{V}_i)\mathbf{J}_i^{G_2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \\
= \;& T(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)^{\mathrm{T}}\mathbf{D}^{G_2}\mathbf{O}_2^{\mathrm{T}}\begin{pmatrix} \widehat{\mathbf{C}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{C}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\mathbf{C}}_m \end{pmatrix}\mathbf{O}_2\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0),
\end{aligned}$$

where $\mathbf{O}_2 = \begin{pmatrix} \mathbf{J}_1^{G_2} \\ \vdots \\ \mathbf{J}_m^{G_2} \end{pmatrix}(\mathbf{D}^{G_2})^{-1}$ is an $mp \times H_2$ orthonormal matrix, and $\widehat{\mathbf{C}}_i = \sum_t(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})^{\mathrm{T}}/T$.

Let $\mathbf{C}_i = E[(g'_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2(\overline{\mathbf{X}}_{it} - E[\overline{\mathbf{X}}_{it}|\mathbf{X}_{it}\boldsymbol{\beta}_{0i}])^{\otimes 2}]$. Lemma 5 in the SUPPLEMENTARY MATERIAL, provided in a separated file, shows that $\max_i \|\widehat{\mathbf{C}}_i - \mathbf{C}_i\|_{op} = o_p(1)$. Based on this, we have the first term in (A.7) is bounded below by $CT\|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)\|^2$.

Now consider the second term in (A.7). We have

$$\sum_{i,t}(\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0)(\mathbf{J}_i^{G_2})^{\mathrm{T}}(\mathbf{V}_{it} - \mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})\epsilon_{it}$$

$$\leq \quad \|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\| \cdot \left\| \sum_t \mathbf{O}_2^{\mathrm{T}} \begin{pmatrix} (\mathbf{V}_{1t}-\mathbf{V}_1^{\mathrm{T}}\mathbf{P}_{1t})\epsilon_{1t} \\ \vdots \\ (\mathbf{V}_{mt}-\mathbf{V}_m^{\mathrm{T}}\mathbf{P}_{mt})\epsilon_{mt} \end{pmatrix} \right\|,$$

and with some more detailed analyses we get (A.3), we get

$$E\left[\left\| \sum_t \mathbf{O}_2^{\mathrm{T}} \begin{pmatrix} (\mathbf{V}_{1t}-\mathbf{V}_1^{\mathrm{T}}\mathbf{P}_{1t})\epsilon_{1t} \\ \vdots \\ (\mathbf{V}_{mt}-\mathbf{V}_m^{\mathrm{T}}\mathbf{P}_{mt})\epsilon_{mt} \end{pmatrix} \right\|^2 \right] = O(H_2 T) \tag{A.8}$$

and thus the second term in (A.7) is $O_p(\sqrt{H_2 T}\|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\|)$. The remaining terms in (A.7) can be shown to be of order $o_p(1)$. Summarizing the bounds for different terms in (A.7), we get

$$\|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\|^2 + \|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\| O_p(\sqrt{H_2/T}) + o_p(1/T) \leq 0,$$

which implies $\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\| = \|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\| = O_p(\sqrt{H_2/T})$.

To get asymptotic normality, we define

$$\widetilde{\boldsymbol{\eta}} = \boldsymbol{\eta}_0 + \left(\sum_{i,t} (\mathbf{J}_i^{G_2})^{\mathrm{T}} (\mathbf{V}_{it}-\mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})(\mathbf{V}_{it}^{\mathrm{T}}-\mathbf{P}_{it}^{\mathrm{T}}\mathbf{V}_i)\mathbf{J}_i^{G_2} \right)^{-1} \sum_{i,t} (\mathbf{J}_i^{G_2})^{\mathrm{T}} (\mathbf{V}_{it}-\mathbf{V}_i^{\mathrm{T}}\mathbf{P}_{it})\epsilon_{it}.$$

Then for any unit vector $\mathbf{a}_2 \in R^{H_2}$, we have

$$\begin{aligned} & \mathbf{a}_2^{\mathrm{T}}\mathbf{D}^{G_2}(\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}_0) \\ = \quad & T^{-1}\mathbf{a}_2^{\mathrm{T}} \left(\mathbf{O}_2^{\mathrm{T}} \begin{pmatrix} \widehat{\mathbf{C}}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{C}}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\mathbf{C}}_m \end{pmatrix} \mathbf{O}_2 \right)^{-1} \cdot \\ & \sum_t \mathbf{O}_2^{\mathrm{T}} \begin{pmatrix} (\mathbf{V}_{1t}-\mathbf{V}_1^{\mathrm{T}}\mathbf{P}_{1t})\epsilon_{1t} \\ \vdots \\ (\mathbf{V}_{mt}-\mathbf{V}_m^{\mathrm{T}}\mathbf{P}_{mt})\epsilon_{mt} \end{pmatrix}. \end{aligned}$$

Consider

$$\begin{aligned} b_2 \quad := \quad & T^{-1}\mathbf{a}_2^{\mathrm{T}} \left(\mathbf{O}_2^{\mathrm{T}} \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_m \end{pmatrix} \mathbf{O}_2 \right)^{-1} \cdot \\ & \sum_t \mathbf{O}_2^{\mathrm{T}} \begin{pmatrix} (\mathbf{V}_{1t}-\boldsymbol{\Phi}_{1t})\epsilon_{1t} \\ \vdots \\ (\mathbf{V}_{mt}-\boldsymbol{\Phi}_{mt})\epsilon_{mt} \end{pmatrix}. \end{aligned}$$

We can show using the central limit theorem under mixing conditions, for example results in Bardet et al. (2008), that

$$\sqrt{T}\nu_{2,T}^{-1/2}b_2 \xrightarrow{d} N(0,1),$$

where

$$\begin{aligned}
\nu_{2,T} &= \mathbf{a}_2^{\mathrm{T}}(\mathbf{O}_2^{\mathrm{T}}\mathbf{C}\mathbf{O}_2)^{-1}\mathbf{O}_2^{\mathrm{T}}\mathbf{\Sigma}_2\mathbf{O}_2(\mathbf{O}_2^{\mathrm{T}}\mathbf{C}\mathbf{O}_2)^{-1}\mathbf{a}_2, \\
\mathbf{C} &= \begin{pmatrix} \mathbf{C}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_m \end{pmatrix}, \\
\mathbf{\Sigma}_2 &= \frac{1}{T}\sum_{1\le t,t'\le T}\begin{bmatrix} \mathbf{C}_{1,|t-t'|} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{2,|t-t'|} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{C}_{m,|t-t'|} \end{bmatrix},
\end{aligned}$$

where $\mathbf{C}_{i,|t-t'|} = E[\epsilon_{it}\epsilon_{it'}g'_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})g'_{0i}(\mathbf{X}_{it'}^{\mathrm{T}}\boldsymbol{\beta}_{0i})(\overline{\mathbf{X}}_{it}-E[\overline{\mathbf{X}}_{it}|\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}])(\overline{\mathbf{X}}_{it'}-E[\overline{\mathbf{X}}_{it'}|\mathbf{X}_{it'}^{\mathrm{T}}\boldsymbol{\beta}_{0i}])^{\mathrm{T}}]$. Furthermore, it can be shown that

$$|\mathbf{a}_2^{\mathrm{T}}\mathbf{D}^{G_2}(\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)-b_1| = o_p(1/\sqrt{T}),$$

and

$$|\mathbf{a}_2^{\mathrm{T}}\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)-\mathbf{a}_2^{\mathrm{T}}\mathbf{D}^{G_2}(\widetilde{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)| = o_p(1/\sqrt{T}),$$

which established the asymptotic normality of $\widehat{\boldsymbol{\eta}}$.

Since $\widehat{\overline{\boldsymbol{\beta}}}-\overline{\boldsymbol{\beta}}_0 = \mathbf{O}_2\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)$, $\mathbf{b}_2^{\mathrm{T}}(\widehat{\overline{\boldsymbol{\beta}}}-\overline{\boldsymbol{\beta}}_0) = \mathbf{b}_2^{\mathrm{T}}\mathbf{O}_2\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)$ is asymptotically normal. That is, for any unit vector $\mathbf{b}_2 \in R^{mp}$,

$$\sqrt{T}\kappa_{2,T}^{-1/2}\mathbf{b}_2^{\mathrm{T}}(\widehat{\overline{\boldsymbol{\beta}}}-\overline{\boldsymbol{\beta}}_0) \xrightarrow{d} N(0,1), \tag{A.9}$$

where

$$\kappa_{2,T} := \mathbf{b}_2^{\mathrm{T}}\mathbf{O}_2(\mathbf{O}_2^{\mathrm{T}}\mathbf{C}\mathbf{O}_2)^{-1}\mathbf{O}_2^{\mathrm{T}}\mathbf{\Sigma}_2\mathbf{O}_2(\mathbf{O}_2^{\mathrm{T}}\mathbf{C}\mathbf{O}_2)^{-1}\mathbf{O}_2^{\mathrm{T}}\mathbf{b}_2.$$

STEP 3. Proof of the convergence rate of $\widehat{\boldsymbol{\theta}}$ and its asymptotic normality.

To get the convergence rate of $\widehat{\boldsymbol{\theta}}$, like for $\widehat{\boldsymbol{\beta}}$, we perform a projection, which is now the projection for the nonparametric part. Let $\mathbf{A}_{0i} := \arg\min_{\mathbf{A}} \|\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g'_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\mathbf{A}\overline{\mathbf{X}}_{it}\|^2$. Obviously, we have

$$\mathbf{A}_{0i} = E\left[g'_{0i}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\overline{\mathbf{X}}_{it}^{\mathrm{T}}\right]\left(E\left[(g'_{0i}(\overline{\mathbf{X}}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}))^2\overline{\mathbf{X}}_{it}\overline{\mathbf{X}}_{it}^{\mathrm{T}}\right]\right)^{-1}.$$

Writing now that

$$\sum_{i,t}(y_{it}-\boldsymbol{\theta}_i^{\mathrm{T}}\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_i))^2$$

$$= \sum_{i,t}(\epsilon_{it} - (\mathbf{\Pi}_{it}^{\mathrm{T}} - \mathbf{Q}_{it}^{\mathrm{T}}\mathbf{\Pi}_i)(\boldsymbol{\theta}_i - \boldsymbol{\theta}_{0i}) - \mathbf{V}_{it}^{\mathrm{T}}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_{0i}) - R_{it1} - R_{it2}(\boldsymbol{\theta}_i, \boldsymbol{\gamma}_i - (\mathbf{V}_i^{\mathrm{T}}\mathbf{V}_i)^{-1}\mathbf{V}_i^{\mathrm{T}}\boldsymbol{\theta}_i))^2,$$

where $\mathbf{Q}_{it}^{\mathrm{T}}$ is the t-th row of $\mathbf{Q}_i = \mathbf{V}_i(\mathbf{V}_i^{\mathrm{T}}\mathbf{V}_i)^{-1}\mathbf{V}_i^{\mathrm{T}}$ and $\boldsymbol{\gamma}_i = \overline{\boldsymbol{\beta}}_i + (\mathbf{V}_i^{\mathrm{T}}\mathbf{V}_i)^{-1}\mathbf{V}_i^{\mathrm{T}}\mathbf{\Pi}_i\boldsymbol{\theta}_i$, $\boldsymbol{\gamma}_{0i} = \overline{\boldsymbol{\beta}}_{0i} + (\mathbf{V}_i^{\mathrm{T}}\mathbf{V}_i)^{-1}\mathbf{V}_i^{\mathrm{T}}\mathbf{\Pi}_i\boldsymbol{\theta}_{0i}$, we can show $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|^2 = O_p(H_1/T)$ and its asymptotic normality similar to arguments used for $\widehat{\boldsymbol{\beta}}$. In particular, we have that for any unit vector $\mathbf{b}_1 \in R^{mp}$,

$$\sqrt{T}\kappa_{1,T}^{-1/2}\mathbf{b}_1^{\mathrm{T}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0,1), \tag{A.10}$$

where

$$\begin{aligned}
\kappa_{1,T} &= \mathbf{b}_1^{\mathrm{T}}\mathbf{O}_1(\mathbf{O}_1^{\mathrm{T}}\mathbf{D}\mathbf{O}_1)^{-1}\mathbf{O}_1^{\mathrm{T}}\mathbf{\Sigma}_1\mathbf{O}_1(\mathbf{O}_1^{\mathrm{T}}\mathbf{D}\mathbf{O}_1)^{-1}\mathbf{O}_1^{\mathrm{T}}\mathbf{b}_1, \\
\mathbf{D} &= \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_m \end{pmatrix}, \\
\mathbf{\Sigma}_1 &= \frac{1}{T}\sum_{1\le t,t'\le T} \begin{bmatrix} \mathbf{D}_{1,|t-t'|} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{2,|t-t'|} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{D}_{m,|t-t'|} \end{bmatrix}, \\
\mathbf{D}_i &= E[(\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g_{0i}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\mathbf{A}_{0i}\overline{\mathbf{X}}_{it})^{\otimes 2}], \\
\mathbf{D}_{i,|t-t'|} &= E[\epsilon_{it}\epsilon_{it'}(\mathbf{B}(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g_{0i}'(\mathbf{X}_{it}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\mathbf{A}_{0i}\overline{\mathbf{X}}_{it})(\mathbf{B}(\mathbf{X}_{it'}^{\mathrm{T}}\boldsymbol{\beta}_{0i}) - g_{0i}'(\mathbf{X}_{it'}^{\mathrm{T}}\boldsymbol{\beta}_{0i})\mathbf{A}_{0i}\overline{\mathbf{X}}_{it'})^{\mathrm{T}}].
\end{aligned}$$

A.4 Proof of Theorems 1 and 2

We now consider the proof of Theorems 1 and 2 as special cases of (A.9) and (A.10). Consider first Theorem 2, under the additional assumption that the true partition is used. As shown previously, the asymptotic variance of $\widehat{\overline{\boldsymbol{\beta}}} - \overline{\boldsymbol{\beta}}_0$ is $T^{-1}\mathbf{O}_2\mathbf{\Theta}_2\mathbf{O}_2^{\mathrm{T}}$, where $\mathbf{\Theta}_2 = (\mathbf{O}_2^{\mathrm{T}}\mathbf{C}\mathbf{O}_2)^{-1}\mathbf{O}_2^{\mathrm{T}}\mathbf{\Sigma}_2\mathbf{O}_2(\mathbf{O}_2^{\mathrm{T}}\mathbf{D}\mathbf{O}_2)^{-1}$. From our proof, it is easy to see that eigenvalues of $\mathbf{\Theta}_2$ are bounded and bounded away from zero. By the definition of the $mp \times H_2$ matrix $\mathbf{O}_2$, it is easy to see that its row corresponding to β_{ij}, say denoted by $\mathbf{O}_{2(ij)}^{\mathrm{T}}$, has a single nonzero entry $1/\sqrt{m_{ij}}$. Let $\mathbf{e}_{ij} = \sqrt{m_{ij}}\mathbf{O}_{2(ij)}$, which is a unit vector, then the asymptotic variance of $\widehat{\beta}_{ij} - \beta_{0ij}$ is $(m_{ij}T)^{-1}\mathbf{e}_{ij}^{\mathrm{T}}\mathbf{\Theta}_2\mathbf{e}_{ij}$.

The asymptotic variance of $\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_{0i}$ is $T^{-1}\mathbf{J}_i^{G_1}(\mathbf{D}^{G_1})^{-1}\overline{\mathbf{\Theta}}_1(\mathbf{D}^{G_1})^{-1}(\mathbf{J}_i^{G_1})^{\mathrm{T}}$, where

$$\overline{\mathbf{\Theta}}_1 = (\mathbf{O}_1^{\mathrm{T}}\mathbf{D}\mathbf{O}_1)^{-1}\mathbf{O}_1^{\mathrm{T}}\mathbf{\Sigma}_1\mathbf{O}_1(\mathbf{O}_1^{\mathrm{T}}\mathbf{C}\mathbf{O}_1)^{-1}$$

with eigenvalues bounded and bounded away from zero. By definition of $\mathbf{J}_i^{G_1}$ and $\mathbf{D}^{G_1}$, it can be seen that each row of the $K \times H_1$ matrix $\mathbf{J}_i^{G_1}(\mathbf{D}^{G_1})^{-1}$ has a single nonzero entry $1/\sqrt{m_i}$ and thus if we define $\mathbf{K}_i = \sqrt{m_i}\mathbf{J}_i^{G_1}(\mathbf{D}^{G_1})^{-1}$, it is easy to directly verify that $\mathbf{K}_i^{\mathrm{T}}\mathbf{v}$ is bounded and bounded away from zero and infinity for any unit vector $\mathbf{v}$. Also, we have $\|\mathbf{B}(x)\| \asymp K$. Thus the asymptotic variance of $\mathbf{B}^{\mathrm{T}}(x)\widehat{\boldsymbol{\theta}}_i - \mathbf{B}^{\mathrm{T}}(x)\boldsymbol{\theta}_{0i}$ can be written as $\frac{K}{m_i T}\mathbf{b}^{\mathrm{T}}(x)\mathbf{\Theta}_1^{\mathrm{T}}\mathbf{b}(x)$, if we define $\mathbf{b}(x) = \mathbf{K}_i^{\mathrm{T}}\mathbf{B}(x)/\|\mathbf{K}_i^{\mathrm{T}}\mathbf{B}(x)\|$, and $\mathbf{\Theta}_1 = \overline{\mathbf{\Theta}}_1\|\mathbf{K}_i^{\mathrm{T}}\mathbf{B}(x)\|^2/K$.

For Theorem 1, since the result is standard, and also is a special case of Theorem 2, we omit the repetition of arguments above. The quantities $\widetilde{\mathbf{e}}_{ij}$, $\widetilde{\mathbf{b}}(x)$, $\widetilde{\boldsymbol{\Theta}}_1$ and $\widetilde{\boldsymbol{\Theta}}_2$ are defined as above based on the trivial structure in which each single parameter forms its own group in the partition.

The proof of Theorem 2 would be complete if we can establish consistency of homogeneity pursuit based on change point detection. That is, we need to show that the true partition can be identified with probability approaching one.

First, we can show $\|\mathbf{D}^{G_2}(\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0)\|_\infty^2 = O_p(\log(Tm)/T)$ and $\|\mathbf{D}^{G_1}(\widehat{\boldsymbol{\xi}}-\boldsymbol{\xi}_0)\|_\infty^2 = O_p(\log(Tm)/T)$. The general strategy for establishing these is similar to showing the convergence rate of $\widehat{\boldsymbol{\beta}}$, using a slightly different projection, and one needs to carefully construct bounds that are valid uniformly over components of $\widehat{\boldsymbol{\eta}}$.

Then we use sequence $b_{(1)} \le \cdots \le b_{(n)}$ $(n = mp)$ for illustration, with estimated change points $\hat{k}_0 = 0 < \hat{k}_1 < \cdots < \hat{k}_{\hat{H}_2} = n$. The true ordered sequence of $\boldsymbol{\beta}$ is $\beta_{0(1)} \le \cdots \le \beta_{0(n)}$ with change points k_h, $h = 0, \ldots, H_2$. Let $\gamma_2 = \min_{2\le h\le H_2} |\beta_{0(k_{h+1})} - \beta_{0(k_h)}|$ be the minimum jump size. The sup-norm convergence results established above, when specializing to the estimator in stage 1, imply that $\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty = O_p(a_T)$ where $a_T = \sqrt{\log(Tm)/T}$. On the event $\{\|\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|_\infty \le Ca_T\}$. It is easy to see that

$$\max_{u-1<k<e} |\Delta_{u,e}(k) - \Delta^0_{u,e}(k)| \le a_T. \tag{A.11}$$

where $\Delta^0_{u,e}(k) = \sqrt{\frac{(e-k)(k-u+1)}{e-u+1}} \left| \frac{\sum_{l=k+1}^{e} \beta_{0(l)}}{e-k} - \frac{\sum_{l=u}^{k} \beta_{0(l)}}{k-u+1} \right|$.

Now suppose $u-1$ and e are both change points and there is at least one change point inside $(u-1, e)$. Let $\widehat{k} = \arg\max_{u-1<k<e} \Delta_{u,e}(k)$ and $k_0 = \arg\max_{u-1<k<e} \Delta^0_{u,e}(k)$. We prove consistency by way of contradiction. Suppose $\widehat{k}$ is not one of the true change points. Using some results in Venkatraman (1992) and Cho and Fryzlewicz (2012), we can show that this would lead to $\Delta_{u,e}(k_0) > \Delta_{u,e}(\widehat{k})$ by (A.11), a contradiction by the definition of $\widehat{k}$. Also, in this case, it is easy to see that $\max_{u-1<k<e} \Delta_{u,e}(k) \ge \max_{u-1<k<e} \Delta^0_{u,e}(k) - a_n \ge C\gamma_2 - a_T > \delta_2$.

Now suppose still $u-1, e$ are both change points but there are no other change point inside $(u-1, e)$. In this case, using (A.11), it is easy to see that $\max_{u-1<k<e} \Delta_{u,e}(k) \le \sqrt{n}a_T$.

Since we refrain from further partitioning the interval $(u-1, e)$ if and only if $\max_{u-1<k<e} \Delta_{u,e}(k) < \delta_2$ with $\sqrt{n}a_T << \delta_2 << \gamma_2$, we see that the algorithm consistently identifies exactly the true change points in $\boldsymbol{\beta}_0$.

The proof for change point detection in $\boldsymbol{\theta}$ is the same, and the proof of Theorem 2 is complete.

A.5 Proof of Theorem 3

For the first statement, we just need to note that $\bar{\boldsymbol{\beta}}$ is the minimizer of

$$\min_{\mathbf{a}} \sum_{i=1}^{m} \|\boldsymbol{\beta}_i - \mathbf{a}\|^2,$$

and all $\check{\boldsymbol{\beta}}_i$ are the same, thus

$$\frac{1}{mp}\sum_{i=1}^{m}\|\check{\boldsymbol{\beta}}_i-\boldsymbol{\beta}_i\|^2 \geq \frac{1}{mp}\sum_{i=1}^{m}\|\boldsymbol{\beta}_i-\bar{\boldsymbol{\beta}}\|^2 \geq c.$$

Similarly we can show the second statement.

References

Bardet, J.-M., Doukhan, P., Lang, G. and Ragache, N. (2008). Dependent lindeberg central limit theorem and some applications, *ESAIM: Probability and Statistics* **12**: 154–172.

Belloni, A., Chernozhukov, V. and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls, *The Review of Economic Studies* **81**(2): 608–650.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR, *Biometrics* **64**(1): 115–123.

Bonhomme, S., Lamadon, T. and Manresa, E. (2017). Discretizing unobserved heterogeneity, *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2019-16).

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data, *Econometrica* **83**(3): 1147–1184.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models, *Journal of the American Statistical Association* **92**: 477–489.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series, *Statistica Sinica* **22**(1): 207–229.

Duan, N. and Li, K. (1991). Slicing regression: a link-free regression method, *The Annals of Statistics* **19**(2): 505–530.

Fan, J., Feng, Y. and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models, *Journal of the American Statistical Association* **106**(494): 544–557.

Fan, J. and Yao, Q. (2003). *Nonlinear time series: nonparametric and parametric methods*, Springer Verlag.

Friedman, J., Hastie, T., Hfling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics* **1**(2): 302–332.

Guo, S., Box, J. L. and Zhang, W. (2017). A dynamic structure for high dimensional covariance matrices and its application in portfolio allocation, *Journal of the American Statistical Association* **112**(517): 235–253.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives, *Journal of the American statistical Association* **84**(408): 986–995.

Hsiao, C. (2014). *Analysis of panel data*, Cambridge university press.

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

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Application in R*, Springer.

Jiang, Q., Wang, H., Xia, Y. and Jiang, G. (2013). On a principal varying coefficient model, *Journal of the American Statistical Association* **108**(501): 228–236.

Ke, Y., Li, J. and Zhang, W. (2016). Structure identification in panel data analysis, *The Annals of Statistics* **44**(3): 1193–1233.

Ke, Z. T., Fan, J. and Wu, Y. (2015). Homogeneity pursuit, *Journal of the American Statistical Association* **110**(509): 175–194.

Kenneth French's website (n.d.).
URL: *http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data%5C_library.html*

Lu, X. and Su, L. (2017). Determining the number of groups in latent panel structures with an application to income and democracy, *Quantitative Economics* **8**: 729–760.

Mirsky, L. (1975). A trace inequality of John von Neumann, *Monatshefte fúr Mathematik* **79**(4): 303–306.

Peng, H. and Huang, T. (2011). Penalized least squares for single index models, *Journal of Statistical Planning and Inference* **141**(4): 1362–1379.

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

UK Met Office website (n.d.).
URL: *http://www.metoffice.gov.uk/public/weather/climate-historic*

van der Geer, S. A. (2000). *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge.

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems*, Thesis, Stanford University.

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

Vogt, M. and Matthias, S. (2017). Clustering with statistical error control, *arxiv: 1702.02643* .

Wang, L. and Yang, L. (2009). Spline estimation of single-index models, *Statistica Sinica* **19**(2): 765–783.

Wang, W., Phillips, P. C. and Su, L. (2018). Homogeneity pursuit in panel data models: Theory and application, *Journal of Applied Econometrics* **33**: 797–825.

Wang, W. and Su, L. (2019). Identifying latent group structures in nonlinear models, *Working paper, Singapore Management University* .

Xia, Y. (2008). A multiple-index model and dimension reduction, *Journal of the American Statistical Association* **103**(484): 1631–1640.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models, *Journal of the American Statistical Association* **97**(460): 1042–1054.

Zhu, L., Miao, B. and Peng, H. (2012). On sliced inverse regression with high-dimensional covariates, *Journal of the American Statistical Association* **101**(474): 630–643.

Zhu, L. and Xue, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3): 549–570.