White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of York.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/98374/

Version: Published Version

Article:
Ke, Yuan, Li, Jialiang and Zhang, Wenyang (2016) Structure Identification in Panel Data Analysis. Annals of Statistics. pp. 1193-1233. ISSN: 0090-5364

https://doi.org/10.1214/15-AOS1403

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

The Annals of Statistics
2016, Vol. 44, No. 3, 1193–1233
DOI: 10.1214/15-AOS1403
© Institute of Mathematical Statistics, 2016

STRUCTURE IDENTIFICATION IN PANEL DATA ANALYSIS[1]

BY YUAN KE, JIALIANG LI AND WENYANG ZHANG

Princeton University, National University of Singapore and University of York

Panel data analysis is an important topic in statistics and econometrics. In such analysis, it is very common to assume the impact of a covariate on the response variable remains constant across all individuals. While the modelling based on this assumption is reasonable when only the global effect is of interest, in general, it may overlook some individual/subgroup attributes of the true covariate impact. In this paper, we propose a data driven approach to identify the groups in panel data with interactive effects induced by latent variables. It is assumed that the impact of a covariate is the same within each group, but different between the groups. An EM based algorithm is proposed to estimate the unknown parameters, and a binary segmentation based algorithm is proposed to detect the grouping. We then establish asymptotic theories to justify the proposed estimation, grouping method, and the modelling idea. Simulation studies are also conducted to compare the proposed method with the existing approaches, and the results obtained favour our method. Finally, the proposed method is applied to analyse a data set about income dynamics, which leads to some interesting findings.

1. Introduction. Panel data analysis is an important topic in statistics and econometrics [Hsiao (2003), Ahn and Schmidt (1995), Arellano (2003), Baltagi (2005)]. Let y_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$, be a one-dimensional response variable, X_{it} be a p-dimensional covariate. The simplest model for panel data analysis would be

$$y_{it} = \alpha_i + X_{it}^{\mathrm{T}}\boldsymbol{\beta} + e_{it}, \qquad i = 1, \ldots, n, t = 1, \ldots, T, \tag{1.1}$$

where α_i, $i = 1, \ldots, n$, and $\boldsymbol{\beta}$ are unknown parameters to be estimated,

$$E(e_{it}|X_{it}) = 0, \qquad \mathrm{Var}(e_{it}|X_{it}) = \sigma^2, \qquad i = 1, \ldots, n, t = 1, \ldots, T. \tag{1.2}$$

However, it is well noted that the constant conditional variance assumption in (1.2) may not be plausible as e_{it}, $t = 1, \ldots, T$, may be correlated within individual and some unobserved latent factors, which influence the covariate X_{it}, may also have an impact on e_{it}. See Bai and Li (2014) for more detailed discussion. To account

Received May 2015; revised September 2015.

[1]Supported by AcRF Grant R-155-000-152-112 and the Singapore National Research Foundation under its Cooperative Basic Research Grant and administered by the Singapore Ministry of Health's National Medical Research Council (Grant No. NMRC/CBRG/0014/2012).

MSC2010 subject classifications. Primary 62F08; secondary 62F12.

Key words and phrases. Binary segmentation algorithm, homogeneity, EM algorithm, interactive effects, panel data.

for the interactive effects caused by such latent factors, Bai and Li (2014) proposed the following model:

$$\begin{cases} y_{it} = \alpha_i + X_{it}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it}, \\ X_{it} = \boldsymbol{\mu}_i + \Gamma_i \mathbf{f}_t + \boldsymbol{\epsilon}_{it}, \end{cases} \qquad i = 1, \ldots, n, t = 1, \ldots, T, \tag{1.3}$$

where: $\mathbf{f}_t$ is a q dimensional latent factor; α_i, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_i$, $\boldsymbol{\mu}_i$ and Γ_i are unknown parameters; ε_{it} and $\boldsymbol{\epsilon}_{it}$ are random errors.

Model (1.3) is useful for panel data analysis with interaction effects. A noticeable characteristic of (1.3) is that it stipulates the impact of a covariate X_{it} on the response y_{it} to be the same across all individuals. If the goal is to provide a global account of the impact of X_{it} on y_{it}, this model may be sensible. However, this model would miss the individual/subgroup attribute of the impact, which may be very important in many cases. Furthermore, from statistical modelling point of view, (1.3) may suffer from estimation and prediction bias. A simple approach to account for the individual/subgroup attribute of the impact would be the following model:

$$\begin{cases} y_{it} = \alpha_i + X_{it}^{\mathrm{T}}\boldsymbol{\beta}_i + \mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it}, \\ X_{it} = \boldsymbol{\mu}_i + \Gamma_i \mathbf{f}_t + \boldsymbol{\epsilon}_{it}, \end{cases} \qquad i = 1, \ldots, n, t = 1, \ldots, T, \tag{1.4}$$

where: ε_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T$, are i.i.d.; $\boldsymbol{\epsilon}_{it}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, are i.i.d.; $\mathbf{f}_t$, $t = 1, \ldots, T$, are i.i.d.; $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{ip})^{\mathrm{T}}$. In this model, we also assume $\mathbf{f}_t$, ε_{it} and $\boldsymbol{\epsilon}_{it}$, $i = 1, \ldots, n$, $t = 1, \ldots, T$, are independent of each other with

$$E(\varepsilon_{it}) = 0, \qquad \operatorname{var}(\varepsilon_{it}) = \sigma_{1i}^2, \qquad E(\mathbf{f}_t) = \mathbf{0}_q, \qquad \operatorname{cov}(\mathbf{f}_t) = \Sigma_f,$$
$$E(\boldsymbol{\epsilon}_{it}) = \mathbf{0}_p, \qquad \operatorname{cov}(\boldsymbol{\epsilon}_{it}) = \sigma_{2i}^2 I_p,$$

where $\mathbf{0}_p$ is a p dimensional vector with each component being 0 and I_p is an identity matrix of size p.

Although (1.4) takes into account the individual attribute of β_{ij}, it involves too many unknown parameters. This model is not parsimonious and would suffer from inflated estimation variances. Further, (1.4) may overlook the inherent homogeneity among the impacts β_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$. Such homogeneity within grouped subjects is equally important as the individual attribute of β_{ij}. To make the model more parsimonious and at the same time allowing the homogeneity among regression coefficients, we impose the following structural condition on (1.4):

$$\beta_{ij} = \begin{cases} \beta_{0,1}, & \text{when } (i, j) \in A_1, \\ \beta_{0,2}, & \text{when } (i, j) \in A_2, \\ \vdots & \quad \vdots \\ \beta_{0,\mathcal{N}+1}, & \text{when } (i, j) \in A_{\mathcal{N}+1}, \end{cases} \tag{1.5}$$

where $\mathcal{N}$ is fixed and $\{A_k : 1 \le k \le \mathcal{N} + 1\}$ is an unknown partition of $\{(i, j) : 1 \le i \le n; 1 \le j \le p\}$ and to be estimated.

Model (1.4) with the homogeneity condition (1.5) is the model we are going to study in this paper. It is easy to see that model (1.4) is not identifiable. To address this issue, we assume Σ_f is a $q \times q$ identity matrix I_q. For $i = 1, \ldots, n$, $\sigma_{1i}^2, \sigma_{2i}^2, \alpha_i$, $\boldsymbol{\lambda}_i$, $\boldsymbol{\mu}_i$ and Γ_i are unknown parameters. For $k = 1, \ldots, \mathcal{N}+1$, $\mathcal{N}$, $\beta_{0,k}$ and A_k are unknown parameters of interest to be estimated. As the main objective of this paper is to explore the impact of the covariate X_{it} on the response y_{it}, which is what people are most interested in reality, we only focus on the estimation for $\mathcal{N}$, $\beta_{0,k}$ and A_k, though all other unknown parameters can be obtained in the estimation procedure.

Regression under homogeneity condition has been studied by quite a few recent works, for example, Bai (1997), Bai and Li (2012), Bai and Ng (2002), Fred and Jain (2003), Harchaoui and Lévy-Leduc (2010), Shen and Huang (2010), Yang et al. (2012), Zhu, Shen and Pan (2013), Tibshirani et al. (2005), Friedman et al. (2007), Bondell and Reich (2008), Jiang et al. (2013), Ke, Fan and Wu (2015) and the reference therein. However, the methods in these works are all based on penalised likelihood/least squares under the framework of treating homogeneity as a kind of sparsity, and the models addressed by these authors are also different from what we study in this paper. The closest model in the literature to the model addressed in this paper is that in Ke, Fan and Wu (2015). The model in Ke, Fan and Wu (2015) is a special case of our model. In particular, the model in Ke, Fan and Wu (2015) does not include any latent variables. The latent variables substantially increase the estimation complexity, but on the other hand they approximate the reality more closely.

Unlike the existing literature that uses penalised likelihood/least squares, we are going to propose an estimation procedure based on a likelihood method coupled with change point detection and binary segmentation algorithm to estimate $\mathcal{N}$, the partition $\{A_k : 1 \le k \le \mathcal{N}+1\}$ and the unknown parameters $\beta_{0,k}$, $k = 1, \ldots, \mathcal{N}+1$. Simulation studies show that our proposed procedure works very well and outperforms the CARDS proposed in Ke, Fan and Wu (2015) under the same simulation settings.

The rest of the paper is organised as follows. We begin in Section 2 with an estimation procedure for the proposed model. In Section 3, we present the asymptotic properties of the proposed estimation procedure. Because there are no closed forms for the proposed estimators, an EM-type algorithm is provided in Section 4 to implement the proposed estimation procedure. Simulation studies are conducted in Section 5 to show how well the proposed estimation procedure works. In Section 6, we apply the proposed model together with the proposed estimation procedure to analyse a data set about the income dynamics in the USA, which leads to some interesting findings. We leave the technical conditions and theoretical proofs of all theoretical results in Section 8.

2. Estimation procedure. Without loss of generality, we assume ε_{it}, $\mathbf{f}_t$ and $\boldsymbol{\epsilon}_{it}$ follow normal distribution. In fact, the proposed estimation procedure is still applicable even when the normality assumption is violated. The estimation proce-

dure consists of three steps: we first treat all β_{ij} as being different and estimate them by the standard maximum likelihood estimation, we call this step initial estimation; then we apply the standard binary segmentation algorithm to detect the homogeneity among β_{ij}; finally, we incorporate the detected homogeneity among β_{ij} into the maximum likelihood estimation to obtain the final estimators of the unknown parameters, we call this step final estimation.

Step 1: *Initial estimation*. Let

$$Z_{it} = (y_{it}, X_{it}^{\mathrm{T}})^{\mathrm{T}}, \qquad \boldsymbol{\gamma}_i = (\alpha_i, \boldsymbol{\mu}_i^{\mathrm{T}})^{\mathrm{T}}, \qquad \mathbf{e}_{it} = (\sigma_{1i}^{-1}\varepsilon_{it}, \sigma_{2i}^{-1}\boldsymbol{\epsilon}_{it}^{\mathrm{T}})^{\mathrm{T}},$$
$$G_i = \begin{pmatrix} 1 & -\boldsymbol{\beta}_i^{\mathrm{T}} \\ \mathbf{0}_p & I_p \end{pmatrix} \quad \text{and} \quad \mathcal{D}_i = \begin{pmatrix} \boldsymbol{\lambda}_i^{\mathrm{T}} \\ \Gamma_i \end{pmatrix}.$$

Then (1.4) can be written as

$$G_i Z_{it} = \boldsymbol{\gamma}_i + \mathcal{D}_i \mathbf{f}_t + \mathbf{e}_{it}, \qquad i = 1, \ldots, n, t = 1, \ldots, T.$$

In matrix form, this is equivalent to

$$\mathbf{G}\mathbf{Z}_t = \mathbf{a} + \mathbf{D}^{\mathrm{T}}\boldsymbol{\xi}_t, \qquad t = 1, \ldots, T, \tag{2.1}$$

where

$$\mathbf{G} = \operatorname{diag}(G_1, \ldots, G_n), \qquad \mathbf{Z}_t = (Z_{1t}^{\mathrm{T}}, \ldots, Z_{nt}^{\mathrm{T}})^{\mathrm{T}}, \qquad \mathbf{a} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\gamma}_n^{\mathrm{T}})^{\mathrm{T}},$$
$$\mathbf{D} = ((\mathcal{D}_1^{\mathrm{T}}, \ldots, \mathcal{D}_n^{\mathrm{T}})^{\mathrm{T}}, \boldsymbol{\Sigma}_{\mathbf{e}}^{1/2})^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{\xi}_t = (\mathbf{f}_t^{\mathrm{T}}, \mathbf{e}_{1t}^{\mathrm{T}}, \ldots, \mathbf{e}_{nt}^{\mathrm{T}})^{\mathrm{T}}.$$

In addition,

$$E(\boldsymbol{\xi}_t) = 0, \qquad \operatorname{cov}(\boldsymbol{\xi}_t) = \boldsymbol{\Sigma},$$

where

$$\boldsymbol{\Sigma}_{\mathbf{e}} = \operatorname{diag}(\Sigma_1, \ldots, \Sigma_n), \qquad \boldsymbol{\Sigma} = \operatorname{diag}(\Sigma_f, I_{n(p+1)}) = I_{n(p+1)+q} \quad \text{and}$$
$$\Sigma_i = \begin{pmatrix} \sigma_{1i}^2 & \mathbf{0}_p^{\mathrm{T}} \\ \mathbf{0}_p & \sigma_{2i}^2 I_p \end{pmatrix} \qquad \text{for } i = 1, \ldots, n.$$

It is easy to see that the log likelihood function of the unknown parameters is

$$L(\boldsymbol{\theta}, \mathbf{D}) = -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D})^{-1}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta}) - \frac{T}{2}\log(|\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}|) - \frac{n(p+1)T}{2}\log(2\pi), \tag{2.2}$$

where

$$\mathcal{X}_t = \operatorname{diag}(\mathbf{B}_{1t}, \ldots, \mathbf{B}_{nt}), \qquad \mathbf{B}_{it} = \begin{pmatrix} 1 & X_{it}^{\mathrm{T}} & \mathbf{0}_p^{\mathrm{T}} \\ \mathbf{0}_p & \mathbf{0}_{p\times p} & I_p \end{pmatrix},$$

$\mathbf{0}_{p\times p}$ is a matrix of size p with each entry being 0, and

$$\boldsymbol{\theta} = (\alpha_1, \boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \alpha_n, \boldsymbol{\beta}_n^{\mathrm{T}}, \boldsymbol{\mu}_n^{\mathrm{T}})^{\mathrm{T}}.$$

The maximiser of (2.2) is the estimator of unknown parameters. In particular, the part of the maximiser that corresponds to $\boldsymbol{\beta}_i$ is an initial estimator, to be denoted by $\tilde{\boldsymbol{\beta}}_i$. Furthermore, we denote the jth entry of $\tilde{\boldsymbol{\beta}}_i$ as $\tilde{\beta}_{ij}$ and use this component as the initial estimator of β_{ij}.

It is well known that the maximiser of (2.2) does not have a closed form [Pinheiro and Bates (2000)]. However, if $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$ were known, the maximiser of (2.2) would be a generalised least squares estimator [Kariya and Kurata (2004)] given by

$$\tilde{\boldsymbol{\theta}} = \left\{\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D})^{-1}\mathcal{X}_t\right\}^{-1} \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D})^{-1}\mathbf{Z}_t. \tag{2.3}$$

The difficulty with the maximisation of (2.2) lies at the unknown $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$. We shall appeal the EM algorithm to solve this problem later. If we ignore the correlation for panel data and treat $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$ as an identity, the resulting estimators are in general not efficient. We investigate such estimators in the simulation studies and refer to them as estimates without covariance adjustment. In contrast, our proposed estimators acknowledging the panel data correlation are referred to as estimates with covariance adjustment.

Step 2: *Detection of homogeneity.* We sort the initial estimators $\tilde{\beta}_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, in ascending order, and denote them by

$$\tilde{\beta}_{(1)} \le \cdots \le \tilde{\beta}_{(np)}.$$

We use R_{ij} to denote the rank of $\tilde{\beta}_{ij}$. The detection of homogeneity is equivalent to the detection of change points among $\tilde{\beta}_{(l)}$, $l = 1, \ldots, np$. To this end, we apply the binary segmentation algorithm in the following.

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

$$S_{i,j}(\kappa) = \frac{1}{j-i}\left\{\sum_{l=i}^{\kappa}(\tilde{\beta}_{(l)} - \bar{\beta}_{i,\kappa})^2 + \sum_{l=\kappa+1}^{j}(\tilde{\beta}_{(l)} - \bar{\beta}_{\kappa+1,j})^2\right\},$$
$$\text{where } \bar{\beta}_{i,\kappa} = \frac{1}{\kappa-i+1}\sum_{l=i}^{\kappa}\tilde{\beta}_{(l)} \text{ and } \bar{\beta}_{\kappa+1,j} = \frac{1}{j-\kappa}\sum_{l=1+\kappa}^{j}\tilde{\beta}_{(l)}. \tag{2.4}$$

Given a threshold δ, which we will elaborate how to select later, the procedure using the binary segmentation algorithm to detect the change points is introduced as follows:

(1) Compute $S_{1,np}(np)$. If $S_{1,np}(np) \le \delta$, there is no change point among $\tilde{\beta}_{(l)}$, $l = 1, \ldots, np$, and the process of detection ends. If $S_{1,np}(np) > \delta$, we minimise $S_{1,np}(\kappa)$ with respect to κ in the range $1, \ldots, np-1$, that is to find $\hat{k}_1$ such that

$$S_{1,np}(\hat{k}_1) = \min_{1 \le \kappa < np} S_{1,np}(\kappa).$$

Then we add $\hat{k}_1$ to the set of change points and divide the region $\{\kappa : 1 \le \kappa \le np\}$ into two subregions: $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$ and $\{\kappa : \hat{k}_1 + 1 \le \kappa \le np\}$.

(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. Compute $S_{1,\hat{k}_1}(\hat{k}_1)$. If $S_{1,\hat{k}_1}(\hat{k}_1) \le \delta$, there is no change point in the region $\{\kappa : 1 \le \kappa \le \hat{k}_1\}$. Otherwise, find $\hat{k}_2$ such that

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

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 $\{\kappa : \hat{k}_1 + 1 \le \kappa \le np\}$, we compute $S_{\hat{k}_1+1,np}(np)$. If $S_{\hat{k}_1+1,np}(np) \le \delta$, there is no change point in the region $\{\kappa : \hat{k}_1 + 1 \le \kappa \le np\}$. Otherwise, find $\hat{k}_3$ such that

$$S_{\hat{k}_1+1,np}(\hat{k}_3) = \min_{\hat{k}_1+1 \le \kappa < np} S_{\hat{k}_1+1,np}(\kappa).$$

Add $\hat{k}_3$ to the set of change points and divide the region $\{\kappa : \hat{k}_1 + 1 \le \kappa \le np\}$ into two subregions: $\{\kappa : \hat{k}_1 + 1 \le \kappa \le \hat{k}_3\}$ and $\{\kappa : \hat{k}_3 + 1 \le \kappa \le np\}$.

(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 $\{\kappa : \hat{k}_1 + 1 \le \kappa \le np\}$ 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

$$\hat{k}_{(1)} < \hat{k}_{(2)} < \cdots < \hat{k}_{(\hat{\mathcal{N}})},$$

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

We use $\hat{\mathcal{N}}$ to estimate $\mathcal{N}$ and

$$\big\{\{(i, j) : \hat{k}_{(s-1)} < R_{ij} \le \hat{k}_{(s)}\} : 1 \le s \le \hat{\mathcal{N}} + 1\big\},$$

to estimate the partition $\{A_s : 1 \le s \le \mathcal{N} + 1\}$. We consider that all β_{ij}s in the same estimated partition have the same value.

Step 3: *Final estimation*. Making use of the homogeneity detected in step 2, we re-parameterise β_{ij} in (1.4) by setting $\beta_{ij} = \zeta_s$ if $\hat{k}_{(s-1)} < R_{ij} \le \hat{k}_{(s)}$ for some s, $1 \le s \le \hat{\mathcal{N}} + 1$. Through this re-parameterisation, the np unknown parameters β_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, are reduced to $\hat{\mathcal{N}} + 1$ unknown parameters ζ_s, $s = 1, \ldots, \hat{\mathcal{N}} + 1$.

Replacing each β_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, in (1.4) by its corresponding ζ_s, we then apply the maximum likelihood estimation to compute the estimator $\hat{\zeta}_s$ of ζ_s, $s = 1, \ldots, \hat{\mathcal{N}} + 1$, and construct the final estimator $\hat{\beta}_{ij}$ of β_{ij} through $\hat{\beta}_{ij} = \hat{\zeta}_s$ if $\hat{k}_{(s-1)} < R_{ij} \le \hat{k}_{(s)}$ for some s, $1 \le s \le \hat{\mathcal{N}} + 1$. The final estimator is $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{ip})^{\mathrm{T}}$.

Apparently, the whole estimation procedure depends crucially on the threshold δ used in step 2. If δ is too small, we would come up with too many groups, leading to inflated variances of the final estimators. On the other hand, if δ is too large, we would mistakenly group different β_{ij}s in the same region and treat them as the same, leading to a biased final estimators. There are many ways to choose δ in practice. In this paper, we use the standard Bayesian information criterion (BIC) to select δ. Our simulation results show the BIC works very well.

3. Asymptotic results. In this section, we are going to provide the asymptotic properties of the estimator of $\boldsymbol{\beta}_i$ constructed in three different cases:

(1) *Overfitting case*, that is the estimator of $\boldsymbol{\beta}_i$ is constructed without using the homogeneity condition (1.5), which is to use the initial estimator, obtained in step 1 of the proposed estimation procedure, as the final estimator. In this case, we denote the estimator of $\boldsymbol{\beta}_i$ by $\tilde{\boldsymbol{\beta}}_i$.

(2) *Correct fitting case*, that is the homogeneity condition (1.5) is taken into account in the construction of the estimator $\hat{\boldsymbol{\beta}}_i$ of $\boldsymbol{\beta}_i$, and the estimator is constructed by the proposed estimation procedure. This is the right approach to the model.

(3) *Mis-specification case*, that is the estimator of $\boldsymbol{\beta}_i$ is constructed under the assumption $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n = \boldsymbol{\beta}^*$ when this assumption is wrong. We use $\check{\boldsymbol{\beta}}$ to denote the resulting estimator. See Bai and Li (2014) for estimation details for this case.

The technical conditions needed for the asymptotic properties stated in this section are introduced in Section 8.1. The proofs of the theorems are given in Section 8.2. In Section 8.3, we provide some technical lemmas to show the uniform consistency of the initial estimators, which are needed in the proofs of the following theorems.

In this paper, we use $\boldsymbol{\beta}^0 = (\boldsymbol{\beta}_1^{0\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{0\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{D}^0$ to denote the true values of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{D}$, respectively. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_n^{\mathrm{T}})^{\mathrm{T}}$ and $\widehat{\boldsymbol{\beta}}^{\mathrm{oracle}}$ be the oracle maximum likelihood estimator of $\boldsymbol{\beta}$ obtained under the assumption that the homogeneity structure was given. For each i, $i = 1, \ldots, n$, let Ω_i be the selection matrix such that $\boldsymbol{\beta}_i = \Omega_i \boldsymbol{\theta}$. Furthermore, we denote $\boldsymbol{\Omega} = (\Omega_1, \ldots, \Omega_n)^{\mathrm{T}}$.

THEOREM 1 (Overfitting case). *Under Conditions 1–2 in Section 8.1, if* $n^2/T \to 0$, *we have*

$$\sqrt{T}\big(\mathbf{\Sigma}_i^*\big)^{-1/2}\big(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\big) \xrightarrow{\mathrm{D}} \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \qquad i = 1, \ldots, n,$$

where

$$\mathbf{\Sigma}_i^* = \lim_{T\to\infty} \Omega_i \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}} \big(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\big)^{-1} \mathcal{X}_t \right)^{-1} \Omega_i^{\mathrm{T}}.$$

The above theorem provides the asymptotic normality of $\tilde{\boldsymbol{\beta}}_i$ which is obtained without using the homogeneity condition (1.5). According to Lemma 3 in Section 8.3, we can see $\tilde{\boldsymbol{\beta}}_i$ is also uniformly consistent with respect to i over $1, \ldots, n$, and the convergence rate is of order $1/\sqrt{T}$.

In the following theorem, with probability approaching one, we show the proposed procedure can correctly detect the homogeneity structure and the convergence rate of the final estimator is of order $1/\sqrt{nT}$, much faster than $1/\sqrt{T}$.

Before presenting the next theorem, we introduce some notation. Let $\mathcal{M}_A$ be a subspace of $\mathbb{R}^{np}$ defined by

$$\mathcal{M}_A = \big\{\boldsymbol{\beta} \in \mathbb{R}^{np} : \beta_{ij} = \beta_{kl}, \text{for any } (i, j), (k, l) \in A_s, 1 \le s \le \mathcal{N} + 1\big\},$$

Ψ_A be a $(\mathcal{N} + 1) \times np$ matrix such that $\boldsymbol{\beta}_A = \Psi_A \boldsymbol{\beta}$, where $\boldsymbol{\beta}_A$ is an $(\mathcal{N} + 1) \times 1$ vector with its sth component $\beta_{A,s}$ being the common coefficient in group A_s. $\boldsymbol{\beta}_A^0$, $\widehat{\boldsymbol{\beta}}_A$ and $\widehat{\boldsymbol{\beta}}_A^{\text{oracle}}$ are defined in the same way.

It is straightforward to see the (s, l)th, $s = 1, \ldots, \mathcal{N} + 1, l = 1, \ldots, np$, entry of Ψ_A is the indicator function $\frac{1}{|A_s|}\mathbf{1}(\beta_l \in A_s)$, where $|A_s|$ is the size of group A_s.

THEOREM 2 (Correct fitting case). *Under Conditions 1–3 in Section 8.1,*

$$P(\hat{\mathcal{N}} = \mathcal{N}) \longrightarrow 1 \quad \textit{and} \quad P\big(\hat{k}_{(s)} = k_{(s)}^0\big) \longrightarrow 1, \qquad s = 1, \ldots, \mathcal{N}.$$

Also, with probability approaching one, the final estimator $\widehat{\boldsymbol{\beta}}$ *is equal to the oracle estimator* $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$. *Furthermore,*

$$\sqrt{nT}\big(\mathbf{\Sigma}_A\big)^{-1/2}\big(\widehat{\boldsymbol{\beta}}_A^{\text{oracle}} - \boldsymbol{\beta}_A^0\big) \xrightarrow{\mathrm{D}} \mathcal{N}(\mathbf{0}, \mathbf{I}_{\mathcal{N}+1}),$$

where

$$\mathbf{\Sigma}_A = \lim_{T\to\infty} \Psi_A \mathbf{\Omega} \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}} \big(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\big)^{-1} \mathcal{X}_t \right)^{-1} \mathbf{\Omega}^{\mathrm{T}} \Psi_A^{\mathrm{T}}.$$

REMARK. We can envisage asymptotic results similar to Theorem 2 that could be derived for the case where p is diverging. However, the technical conditions would be much stronger, and the technical proofs would be much more involved.

Now we consider the asymptotic behaviour of the estimator in the mis-specification case. We denote $\check{\boldsymbol{\beta}} = (\check{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \check{\boldsymbol{\beta}}_n^{\mathrm{T}})^{\mathrm{T}}$ the maximum likelihood estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_n^{\mathrm{T}})^{\mathrm{T}}$ based on the mis-specified model which assumes $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n = \boldsymbol{\beta}^*$. In the following theorem, we will show $\check{\boldsymbol{\beta}}$ is inconsistent. Therefore, the detection of homogeneity is necessary as it will not only reduce the variance compared with the over-fitting case, but also avoid inconsistent estimation compared with the mis-specification case.

THEOREM 3 (Mis-specification case). *Under Conditions 1–4 in Section 8.1, with probability* 1 *we have*

$$\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| \geq \mathcal{C}\sqrt{n},$$

where $\mathcal{C} > 0$.

4. Computational algorithm. Regarding the computational issues, the main difficulty to implement the proposed estimation procedure is the maximisation of the log likelihood function of unknown parameters. One can see that the maximisation of the log likelihood function in the final estimation step is the same as that in the initial estimation step. Therefore, we only introduce the maximisation in the initial estimation step, that is the maximisation of (2.2).

Let

$$\mathbf{H} = \left(\mathcal{D}_1^{\mathrm{T}}, \ldots, \mathcal{D}_n^{\mathrm{T}}\right)^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{\Sigma}_0 = \operatorname{diag}(\Sigma_1, \ldots, \Sigma_n).$$

It is easy to see

$$\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D} = \mathbf{H}\Sigma_f\mathbf{H}^{\mathrm{T}} + \boldsymbol{\Sigma}_0.$$

We apply the following iterative procedure to maximise (2.2):

(1) In the first iteration, we set the initial values of $\mathbf{H}$ and $\boldsymbol{\Sigma}_0$ such that $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$ is $I_{n(p+1)}$. By (2.3), we have the initial value of $\boldsymbol{\theta}$, which is

$$\boldsymbol{\theta}_{(0)} = \left\{\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\right\}^{-1} \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\mathbf{Z}_t.$$

We fix the $\boldsymbol{\theta}$ in (1.4) at $\boldsymbol{\theta}_{(0)}$. Then we apply the EM algorithm, which will be detailed later, to update $\mathbf{H}$ and $\boldsymbol{\Sigma}_0$ and thereby $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$. The updated $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$ is denoted by $\mathbf{D}_{(1)}^{\mathrm{T}}\boldsymbol{\Sigma}_{(1)}\mathbf{D}_{(1)}$. Replacing the $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$ in (2.3) by $\mathbf{D}_{(1)}^{\mathrm{T}}\boldsymbol{\Sigma}_{(1)}\mathbf{D}_{(1)}$, we can update $\boldsymbol{\theta}$ to

$$\boldsymbol{\theta}_{(1)} = \left\{\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\left(\mathbf{D}_{(1)}^{\mathrm{T}}\boldsymbol{\Sigma}_{(1)}\mathbf{D}_{(1)}\right)^{-1}\mathcal{X}_t\right\}^{-1} \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\left(\mathbf{D}_{(1)}^{\mathrm{T}}\boldsymbol{\Sigma}_{(1)}\mathbf{D}_{(1)}\right)^{-1}\mathbf{Z}_t.$$

(2) In the second iteration, we fix the $\boldsymbol{\theta}$ in (1.4) at $\boldsymbol{\theta}_{(1)}$. By applying the EM algorithm, we can get the updated $\mathbf{H}$ and $\boldsymbol{\Sigma}_0$ and thereby the updated $\mathbf{D}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}$ which is denoted by $\mathbf{D}_{(2)}^{\mathrm{T}}\boldsymbol{\Sigma}_{(2)}\mathbf{D}_{(2)}$. Again, by (2.3), we can update $\boldsymbol{\theta}$ to $\boldsymbol{\theta}_{(2)}$.

(3) We continue this iterative algorithm until convergence.

In the following part, we are going to introduce the EM algorithm used to update $\mathbf{H}$ and $\boldsymbol{\Sigma}_0$ in the above iterative procedure.

From model (1.4), one can see the log likelihood function of the unknown parameters based on $\{(\mathbf{Z}_t^{\mathrm{T}}, \mathbf{f}_t^{\mathrm{T}}) : t = 1, \ldots, T\}$ is

$$\begin{aligned} L = {} & -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta} - \mathbf{H}\mathbf{f}_t)^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta} - \mathbf{H}\mathbf{f}_t) - \frac{T}{2}\log\bigl(|\boldsymbol{\Sigma}_0|\bigr) \\ & - \frac{1}{2}\sum_{t=1}^{T}\mathbf{f}_t^{\mathrm{T}}\Sigma_f^{-1}\mathbf{f}_t \\ & - \frac{T}{2}\log\bigl(|\Sigma_f|\bigr) - \frac{qT}{2}\log(2\pi) - \frac{n(p+1)T}{2}\log(2\pi). \end{aligned}$$

By some calculations, we can see the conditional distribution of $\mathbf{f}_t$ given $\mathbf{Z}_t$ is

$$N\bigl((\mathbf{H}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{H} + \Sigma_f^{-1})^{-1}\mathbf{H}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta}), (\mathbf{H}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{H} + \Sigma_f^{-1})^{-1}\bigr).$$

We denote the mean and covariance matrix of the conditional distribution of $\mathbf{f}_t$ given $\mathbf{Z}_t$ after kth iteration by $M_{t,(k)}$ and $\mathbf{V}_{(k)}$ respectively, and the estimated $\boldsymbol{\theta}$ after the kth iteration by $\boldsymbol{\theta}_{(k)}$. Also let

$$\begin{aligned} \mathcal{Z}_{(k)} &\equiv \bigl(\{\mathbf{Z}_1 - \mathcal{X}_1\boldsymbol{\theta}_{(k)}\}^{\mathrm{T}}, \ldots, \{\mathbf{Z}_T - \mathcal{X}_T\boldsymbol{\theta}_{(k)}\}^{\mathrm{T}}\bigr)^{\mathrm{T}}, \\ \mathbf{m}_{(k)} &= (M_{1,(k)}, \ldots, M_{T,(k)})^{\mathrm{T}}. \end{aligned}$$

After the kth iteration, the conditional expectation of L, with respect to the conditional distribution of $\{\mathbf{f}_1, \ldots, \mathbf{f}_T\}$ given $\{\mathbf{Z}_1, \ldots, \mathbf{Z}_T\}$ and $\boldsymbol{\theta}_{(k)}$, is

$$\begin{aligned} & L_{(k)}(\mathbf{H}, \sigma_{11}, \ldots, \sigma_{1n}, \sigma_{21}, \ldots, \sigma_{2n}) \\ & \quad = -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta}_{(k)} - \mathbf{H}M_{t,(k)})^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta}_{(k)} - \mathbf{H}M_{t,(k)}) \\ & \qquad - \frac{T}{2}\operatorname{tr}\bigl(\mathbf{H}^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{H}\mathbf{V}_{(k)}\bigr) \\ & \qquad - \frac{1}{2}\sum_{t=1}^{T}M_{t,(k)}^{\mathrm{T}}\Sigma_f^{-1}M_{t,(k)} - \frac{T}{2}\bigl\{\log\bigl(|\boldsymbol{\Sigma}_0|\bigr) + \operatorname{tr}\bigl(\Sigma_f^{-1}\mathbf{V}_{(k)}\bigr) + \log\bigl(|\Sigma_f|\bigr)\bigr\} \\ & \qquad - \frac{qT}{2}\log(2\pi) - \frac{n(p+1)T}{2}\log(2\pi). \end{aligned}$$

By maximising $L_{(k)}(\mathbf{H}, \sigma_{11}, \ldots, \sigma_{1n}, \sigma_{21}, \ldots, \sigma_{2n})$ with respect to $\mathbf{H}$, σ_{1i} and σ_{2i}, $i = 1, \ldots, n$, we can obtain the maximiser as

$$\mathbf{H}_{(k+1)} = \mathcal{Z}_{(k)}^{\mathrm{T}} \mathbf{m}_{(k)} \big(\mathbf{m}_{(k)}^{\mathrm{T}} \mathbf{m}_{(k)} + T\mathbf{V}_{(k)}\big)^{-1}, \qquad \sigma_{1i,(k+1)}^2 = \frac{1}{T} \mathrm{u}_i^{\mathrm{T}} \mathcal{Z}_{(k)}^{\mathrm{T}} \mathbf{P}_{(k)} \mathcal{Z}_{(k)} \mathrm{u}_i$$

and

$$\sigma_{2i,(k+1)}^2 = \frac{1}{pT} \operatorname{tr}\big(\mathrm{v}_i \mathcal{Z}_{(k)}^{\mathrm{T}} \mathbf{P}_{(k)} \mathcal{Z}_{(k)} \mathrm{v}_i^{\mathrm{T}}\big),$$

where

$$\mathbf{P}_{(k)} = I_T - \mathbf{m}_{(k)} \big(\mathbf{m}_{(k)}^{\mathrm{T}} \mathbf{m}_{(k)} + T\mathbf{V}_{(k)}\big)^{-1} \mathbf{m}_{(k)}^{\mathrm{T}},$$

$$\mathrm{u}_i = \varsigma_{(p+1)(i-1)+1,(p+1)n}, \qquad \mathrm{v}_i = (\mathbf{0}_{p\times((p+1)(i-1)+1)}, I_p, \mathbf{0}_{p\times((p+1)(n-i))}),$$

and $\varsigma_{i,k}$ is a unit vector with length k and ith component being 1.

Then we use $\mathbf{H}_{(k+1)}$, $\sigma_{1i,(k+1)}^2$ and $\sigma_{2i,(k+1)}^2$, $i = 1, \ldots, n$ to update $\mathbf{\Sigma}_0$ in the $(k+1)$th iteration.

5. Simulation studies. In this section, we conduct simulations to assess the performance of the proposed procedure. In Section 5.1, we examine the accuracy of homogeneity detection of our procedure and compare it with some existing approaches. In Section 5.2, we test the estimation accuracy of our procedure for complicated panel data.

5.1. *Accuracy of homogeneity detection.* Most existing homogeneity detection methods are confined to cross-sectional data. One representative is the CARDS method proposed by Ke, Fan and Wu (2015). We compare our procedure with the CARDS in this subsection. To make the comparison convincing, we keep the simulation settings the same as that in Ke, Fan and Wu (2015), and use the same criterion to measure the accuracy of detection. We note that the model considered in the numerical study of Ke, Fan and Wu (2015) is a linear regression model. Nevertheless, the goal of identifying homogeneous regression coefficients is essentially the same as what we study in this paper for factor model.

In the following, we abbreviate the change point detection method proposed in this paper as the CPD method. The detailed simulation settings are as follows.

We generate the sample from

$$y_i = X_i^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_i, \tag{5.1}$$

where $\{X_i : 1 \le i \le n\}$ are independently generated from the standard multivariate normal distribution, and $\{\varepsilon_i : 1 \le i \le n\}$ are independently generated from standard normal distribution and independent with the covariates. All the simulation results are based on 100 replicates.

We set $p = 60$ and $n = 100$. Predictors are divided into four groups with each group of size 15. The four different values of the true regression coefficients are

$-2r$, $-r$, r, and $2r$, respectively. We consider different values of $r > 0$ for various signal-to-noise ratios.

In this example, the BIC we used to select δ in the CPD method is

$$\sum_{i=1}^{n}\bigl(y_i - X_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}\bigr)^2 + \#\log(n), \tag{5.2}$$

where # is the total number of distinct parameters in the estimated model.

We compare the performance of the CPD method with 5 existing methods along with the oracle estimators. Performance is evaluated in terms of the average prediction error over an independent test sample of size 10,000 from model (5.1). In addition, we consider the normalized mutual information (NMI) to measure how close the estimated grouping structure approaches the true structure. Suppose $\mathbb{C} = \{C_1, C_2, \ldots\}$ and $\mathbb{D} = \{D_1, D_2, \ldots\}$ are two sets of disjoint cluster of $\{1, 2, \ldots, p\}$. The NMI is defined as

$$\mathrm{NMI}(\mathbb{C}, \mathbb{D}) = \frac{2I(\mathbb{C}, \mathbb{D})}{H(\mathbb{C}) + H(\mathbb{D})}, \tag{5.3}$$

where

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

is the mutual information between the two clusterings, and

$$H(\mathbb{C}) = -\sum_{k}\bigl(|C_k|/p\bigr)\log\bigl(|C_k|/p\bigr)$$

is the entropy of $\mathbb{C}$. NMI ranges between 0 and 1 with large values indicating a higher degree of similarity between the two clusterings. The results are reported in Table 1. Because we use the same simulation settings as that in Ke, Fan and Wu (2015), the results reported in their paper are adapted in Table 1.

Table 1 shows the CPD method performs very well as it has the largest NMI values across all settings, indicating that it can recover the true grouping structure. The prediction errors are also comparable to CARDS and in general better than TV, fLASSO and OLS.

5.2. *Accuracy of estimation*. We generate a sample from model (1.4) with the homogeneity condition (1.5) and set

$$n = 50, \qquad T = 50, \qquad p = 4, \qquad q = 3, \qquad \Sigma_f = \begin{pmatrix} 1 & 0.75 & 0.75 \\ 0.75 & 1 & 0.75 \\ 0.75 & 0.75 & 1 \end{pmatrix},$$

$\alpha_i = 1$ for all i, β_{i1} being the realization of a sample which is independently drawn from a discrete uniform with atoms $\{-2r, r\}$, β_{i2} being the realization of a sample

TABLE 1
Simulation results for homogeneity detection

r	Oracle	OLS	CPD	bCARDS	aCARDS	TV	fLASSO
			Medians of the average prediction error over 100 repetitions				
1.0	1.0355	1.6112	1.0390	1.0504	1.1182	1.4847	1.4253
0.9	1.0273	1.5885	1.0417	1.0479	1.1048	1.4608	1.4186
0.8	1.0359	1.5947	1.0811	1.0826	1.1786	1.4777	1.4427
0.7	1.0311	1.6038	1.2417	1.1250	1.2830	1.5591	1.4625
0.6	1.0370	1.6054	1.4095	1.3172	1.4586	1.5795	1.4824
0.5	1.0347	1.5826	1.4536	1.3645	1.5734	1.5734	1.4668
			Medians of the NMI over 100 repetitions				
1.0	1.0	0.5059	1.0000	0.9414	0.9784	0.7203	0.6503
0.9	1.0	0.5059	0.9911	0.9414	0.9784	0.7167	0.6521
0.8	1.0	0.5059	0.9762	0.8609	0.9355	0.7245	0.6549
0.7	1.0	0.5059	0.9650	0.7912	0.8989	0.6991	0.6458
0.6	1.0	0.5059	0.9170	0.7008	0.8763	0.6808	0.6373
0.5	1.0	0.5059	0.8575	0.6722	0.6741	0.6654	0.6251

"Oracle" refers to the least squares estimator with knowing the true grouping; "OLS" refers to the ordinary least squares estimator with no grouping; "CPD" refers to our change point detection method; "bCARDS" and "aCARDS" are two versions of clustering algorithm in regression via data-driven segmentation; "TV" refers to the total variation method; "fLASSO" refers to the fused LASSO.

independently drawn from a discrete uniform with atoms $\{-r, 2r\}$, β_{i3} being the realization of a sample which is independently drawn from a discrete uniform with atoms $\{-2r, -r\}$, β_{i4} being the realization of a sample independently drawn from a discrete uniform with atoms $\{r, 2r\}$. The elements of $\boldsymbol{\lambda}_i$, $\boldsymbol{\mu}_i$ and Γ_i are set to be the realizations of samples generated independently from the standard normal distribution. $\mathbf{f}_t$ is generated independently from the multivariate normal distribution with mean zero and covariance matrix Σ_f. ε_{it} and $\boldsymbol{\epsilon}_{it}$ are generated independently from the standard normal distribution. We deliberately set Σ_f non-identity matrix to show our method still works very well for such situation.

To have a deep insight about the advantage of the proposed estimation procedure, for each generated sample, we evaluate the pre-grouping initial estimate without covariance adjustment (E1), the pre-grouping estimate adjusted for covariance (E2), the post-grouping initial estimate without covariance adjustment (E3) and the post-grouping final estimate adjusted for covariance (E4), which is the proposed estimation procedure. After 500 simulations, we summarise the mean squared error and mean absolute errors for regression parameters in Table 2. The simulation results show our final estimates are consistent to the true parameters. We notice that the estimation was drastically improved after we considered the homogeneity of the regression coefficients. The covariance adjusted estimators are generally closer to the true parameters than the naive estimates without covariance adjustment.

TABLE 2
Estimation results over 500 *repetitions*

		μ		α		β	
r	**Estimate**	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**	**MAE**
2.0	E1	10.64	2.92	0.9025	0.6968	10.61	2.92
	E2	10.45	2.91	0.2762	0.3971	10.61	2.92
	E3	0.1917	0.3358	0.1751	0.3246	0.0097	0.0763
	E4	0.1917	0.3358	0.1686	0.3149	0.0062	0.0646
1.0	E1	3.4102	1.5485	0.9165	.6846	3.3414	1.5343
	E2	3.2325	1.5149	0.2655	0.3870	3.3414	1.5343
	E3	0.1573	0.2875	0.2347	0.3374	0.0098	0.0815
	E4	0.1573	0.2875	0.2321	0.3349	0.0065	0.0667
0.75	E1	2.4823	1.2914	0.9622	.7043	2.4852	1.2857
	E2	2.3126	1.2487	0.3100	0.4176	2.4852	1.2857
	E3	0.2072	0.3423	0.3313	0.4183	0.0081	0.0713
	E4	0.2072	0.3423	0.3239	0.4129	0.0079	0.0683

"E1" refers to the pre-grouping initial estimate without covariance adjustment, "E2" refers to the pre-grouping estimate adjusted for covariance, "E3" refers to the post-grouping initial estimate without covariance adjustment, and "E4" refers to the post-grouping final estimate adjusted for covariance. "MSE" is the mean squared error and "MAE" is the mean absolute error, averaged over all elements of the parameters.

6. Real data analysis. In this section, we study a real data of the non-Survey of Economic Opportunity portion of the Panel Study of Income Dynamics (PSID). This data is drawn from 1976 to 1982. Starting with a national sample of 5000 U.S. households in 1968, the PSID re-interviewed individuals from those households over the years. These individuals are re-interviewed whether or not they are living in the same house or with the same people. New households are added to the sample when the children of the panel families grow up and start their own family. The sample size has increased from about 4800 core households in 1968 to almost 10,700 in 1992. The study is conducted by the Survey Research Center, Institute for Social Research, University of Michigan (home page: http://www.isr.umich.edu/src/psid/).

The individuals in our sample are 595 heads of household between the ages of 18 and 65 in 1976, who report a positive wage in some private, non-farm employment for all 7 years. The response variable is the logarithm of wage. The predictors include years of full-time work experience (EXP), the squared term of experience (EXP2), weeks worked (WKS), occupation (OCC = 1, if the individual has blue-collar occupation), industry (IND = 1, if the individual works in a manufacturing industry), residence (SOUTH = 1, SMSA = 1, if the individual resides in the south, or in a standard metropolitan statistical area), marital status (MS = 1,

TABLE 3
Estimation results for PSID analysis

	CPD2		CPD1		LME0		LME1		BL	
	Coef.	**SE**	**Coef.**	**SE**	**Coef.**	**SE**	**Coef.**	**SE**	**Coef.**	**SE**
EXP	0.1124	0.0006	0.1125	0.0163	0.1078	0.0024	0.1072	0.0024	0.1342	0.0217
	0.1388	0.0027	0.1374	0.0324	–	–	0.1257	0.0041	–	–
	0.1740	0.0115	0.1741	0.0965	–	–	0.1437	0.0116	–	–
EXP2	−0.0005	0.00001	−0.0005	0.0004	−0.0005	0.00005	−0.0005	0.00005	−0.0005	0.0004
WKS	0.0005	0.0001	0.0009	0.0040	0.0008	0.0006	0.0009	0.0006	0.0001	0.0042
OCC	−0.0108	0.0034	−0.0220	0.0907	−0.0396	0.0137	−0.0383	0.0136	−0.0001	0.1019
IND	0.0173	0.0039	0.0187	0.1017	0.0088	0.0153	0.0082	0.0152	0.0086	0.1068
SOUTH	0.0003	0.0172	0.0017	0.2257	−0.0161	0.0320	−0.0126	0.0319	0.0702	0.2776
SMSA	−0.0363	0.0080	−0.0345	0.1284	−0.0401	0.0190	−0.0349	0.0189	−0.0437	0.1545
MS	−0.0243	0.0050	−0.0287	0.1249	−0.0354	0.0188	−0.0347	0.0187	−0.2283	0.1612
UNION	0.0182	0.0041	0.0313	0.0983	0.0330	0.0148	0.0333	0.0147	0.0370	0.1050

"CPD2" is our proposed change point detection method with covariance adjustment; "CPD1" is the change point detection method without covariance adjustment; "LME0" is the linear mixed effects model with constant coefficient; "LME1" is the linear mixed effects model with non-constant coefficients; "BL" is the Bai-Li estimator. "Coef." is the estimated regression coefficient and "SE" is the standard error.

if the individual is married) and union coverage (UNION = 1, if the individual's wage is set by a union contract).

We first estimate regression coefficients using model (1.4) and then apply the change point detection algorithm for each predictor variable. The algorithm detected non-constant coefficients for EXP have three different groups and did not detect any non-constant coefficients for all other predictors. The individual estimates of EXP coefficients before and after grouping are plotted in Figure 1. The final estimates are presented in Table 3 along with standard errors. The wage gains from an additional year of past work experience are 0.11, 0.13 and 0.18, respectively, for the three groups of individuals, indicating a positive association. We compared estimates obtained with and without covariance adjustment, denoted by CPD2 and CPD1, respectively, in Table 3. The coefficient estimates are close but the standard errors of CPD1 estimates are much larger than those of CPD2 estimates. Hence, the covariance adjustment can improve the estimation efficiency.

In Table 3, we also compared the regression estimates from the standard linear mixed effects (LME) models without and with grouping, denoted by LME0 and LME1, respectively. When all individuals are placed in a single group, the wage gain for an additional year of past work experience is only 0.1078. This estimated moderate overall effect may result from a dominating class of individuals in the sample and implicitly masks the stronger effects of some subgroups. Interpretation of regression estimates from LME is similar to that of model (1.4). However, LME models do not lend support for the latent factors. Further, we implemented

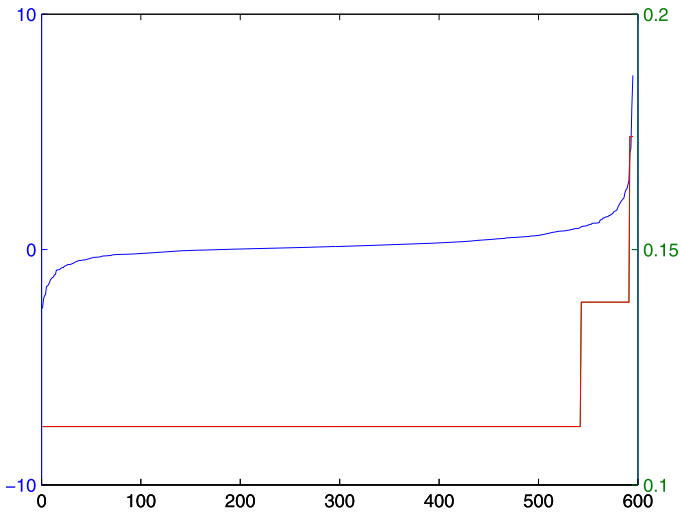

FIG. 1. *Estimated coefficients for EXP for all individuals in PSID data. The blue dashed line is the sorted estimate for individual subjects before grouping. The green solid line is the sorted estimate after grouping. The Web version of this plot is coloured.*

the estimator in Bai and Li (2014) for the PSID data and reported the fitted results in Table 3 as well. Though their estimated coefficients are similar to our estimates, the standard errors of Bai–Li estimators are much larger. Model 1.3 does not incorporate the proper coefficient structure, and consequently the model estimates inflate the estimation variation.

The estimated latent factors f_t are -60.69, -42.94, -23.66, -2.88, 19.28, 42.98 and 67.93 for the years 1976–1982. There is an increasing trend for the mean log wage during this period, roughly 20 wage gain per year for a subject with a unit random effect λ_i. In our model, individuals' random effects cooperate with the latent year effects to generate the time-varying regressors and outcomes. Specifically, comparing two individuals observed at the same year t, the predicted income difference should depend on their own characteristics, say λ_i, λ_j, as well as the contemporary year effect. This is in contrast with the standard LME model where only subject-specific random effects are modelled and the random effects are assumed to be the same across the years. The more sophisticated latent factor structure in our model may help explaining how individuals' unobserved latent characteristics progress in the years and their variation may be fully examined via the latent factor independent of the individual's personal working ability or skill.

The estimated random effects λ_i and Γ_i are also available from our analysis. For space consideration, we report them in the online supplementary file of the paper [Ke, Li and Zhang (2015)]. We have also fitted the model with $q = 2$ and $q = 3$. The results are very similar to what we have reported in this section.

7. Discussion. We have assumed the factors $\mathbf{f}_t$ to be i.i.d. in this paper. Sometimes it may be of interest to consider a more complicated setting with $\operatorname{cov}(\mathbf{f}_t, \mathbf{f}_{t'}) \neq 0$. This setting could be plausible when there are multi-level of unobserved latent factors or the random factors may progress dependently over the time. Our estimation procedure can be modified to satisfy this kind of advanced requirement. Throughout this paper, we fix the number of factors q. A more objective selection method based on information criteria may be adopted to determine an optimal q value. We are still developing the necessary theoretical and computational framework.

8. Technical conditions and proofs of the theoretical results. In Section 8.1, we introduce the technical conditions that are needed to prove the asymptotic results in this paper. In Sections 8.2 and 8.3, we provide the detailed proofs of the main theoretical results and some technical lemmas, respectively.

8.1. *Technical conditions.* Set $\mathbf{V}_t = (Y_t, \mathbf{X}_{1t}^{\mathrm{T}}, \ldots, \mathbf{X}_{nt}^{\mathrm{T}})^{\mathrm{T}}$, $t = 1, \ldots, T$. Let $L(\boldsymbol{\Theta})$ be the log likelihood function of observations $\mathbf{V}_1, \ldots, \mathbf{V}_T$, where $\boldsymbol{\Theta}$ is a vector of all unknown parameters and contains $\boldsymbol{\theta}$ as its sub-vector, that is, every entry of $\boldsymbol{\theta}$ is also an entry of $\boldsymbol{\Theta}$. Then by the model assumption, $\mathbf{V}_1, \ldots, \mathbf{V}_T$ are independent and identical distributed observations with probability density $f(\mathbf{V}, \boldsymbol{\Theta})$ with respect to some measure μ, and $f(\mathbf{V}, \boldsymbol{\Theta})$ has common support. In addition, one can easily see the length of $\boldsymbol{\Theta}$ (i.e., the number of unknown parameters) is $n\{(p+1)(q+2)+1\}$. Denote Ψ the parameter space for $\boldsymbol{\Theta}$.

Furthermore, for $j, k, l = 1, \ldots, n\{(p+1)(q+2)+1\}$, we assume $f(\mathbf{V}, \boldsymbol{\Theta})$ satisfies the following Condition 1:

CONDITION 1. (i)

$$E_{\boldsymbol{\Theta}}\left[\frac{\partial \log f(\mathbf{V}, \boldsymbol{\Theta})}{\partial \Theta_j}\right] = 0.$$

(ii) The Fisher information matrix

$$I(\boldsymbol{\Theta}) = E_{\boldsymbol{\Theta}}\left\{\left[\frac{\partial}{\partial \boldsymbol{\Theta}} \log f(\mathbf{V}, \boldsymbol{\Theta})\right]\left[\frac{\partial}{\partial \boldsymbol{\Theta}} \log f(\mathbf{V}, \boldsymbol{\Theta})\right]^{\mathrm{T}}\right\}$$

is finite and positive definite at $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$.

(iii) There exists an open subset ψ of Ψ that contains the true parameter point $\boldsymbol{\Theta}_0$ such that for almost all $\mathbf{V}$ the density $f(\mathbf{V}, \boldsymbol{\Theta})$ admits all third derivatives $(\partial f(\mathbf{V}, \boldsymbol{\Theta}))/(\partial \Theta_j\, \partial \Theta_k\, \partial \Theta_l)$ for all $\boldsymbol{\Theta} \in \psi$. There exist functions M_{jkl} such that

$$\left|\frac{\partial^3}{\partial \Theta_j\, \partial \Theta_k\, \partial \Theta_l} \log f(\mathbf{V}, \boldsymbol{\Theta})\right| \leq M_{jkl}(\mathbf{V}) \qquad \text{for all } \boldsymbol{\Theta} \in \psi,$$

where $m_{jkl} = E_{\boldsymbol{\Theta}_0}[M_{jkl}(\mathbf{V})] \leq \infty$ for all j, k, l.

CONDITION 2. Denote $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ the smallest and largest eigenvalue of a given matrix, respectively. Suppose C is a sufficiently large positive constant. For all $i = 1, \ldots, n$, and $t = 1, \ldots, T$:

(i) $E(\|\mathbf{f}_t\|_F^4) \le C$, $E(\varepsilon_{it}^4) \le C$, $E(\|\boldsymbol{\epsilon}_{it}\|_F^4) \le C$ and $\max_i \|\boldsymbol{\lambda}_i\| \le C$.

(ii) The covariates satisfy $\max_{i,t} \|X_{it}\| \le C$, where $\max_{i,t}$ means the maximum among all the possible pairs of $\{i, t\}$, and $\lambda_{\min}(\frac{1}{T}\sum_{t=1}^T \mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t) \ge C^{-1}$.

(iii) The true value of the generalised loading matrix, that is, $\mathbf{D}^0$, satisfies

$$C^{-1} \le \lambda_{\min}\big(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\big)^{-1} \le \lambda_{\max}\big(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\big)^{-1} \le C,$$

and $\|\frac{1}{T}\sum_{t=1}^T \mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\|) \le C$.

(iv) The limit $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t$ exists.

CONDITION 3. (i) The number of groups $\mathcal{N}+1$ is fixed. The size of each group satisfies $|A_s| = O_p(np)$ for $s = 1, \ldots, \mathcal{N}+1$.

(ii) The smallest gap between two consecutive groups is lower bounded, that is, for $s = 1, \ldots, \mathcal{N}+1$,

$$\mathcal{J} = \min_s |\beta_{0,s+1} - \beta_{0,s}| > C^{-1}.$$

(iii) Recall (2.4) and assume

$$S_{1,np}\big(k_{i_1}^0\big) < S_{1,np}\big(k_{i_2}^0\big) < \cdots < S_{1,np}\big(k_{i_{\mathcal{N}}}^0\big),$$

where $\{i_1, i_2, \ldots, i_{\mathcal{N}}\}$ is a permutation of $\{1, 2, \ldots, \mathcal{N}\}$.

(iv) $n \log n = o(\sqrt{T})$.

(v) When $np \to \infty$, $\delta \to 0$ and $np\delta \to \infty$.

CONDITION 4. There exists an integer j_0, $1 \le j_0 \le p$, and two positive constants $\mathcal{G}_0$ and $\mathcal{G}_1$, such that

$$0 < \mathcal{G}_0 < \max_{1\le s\le\mathcal{N}+1} \frac{|A_{s,j_0}|}{n} = \mathcal{G}_n < \mathcal{G}_1 < 1,$$

where $A_{s,j_0} = \{i : (i, j_0) \in A_s\}$, $|A_{s,j_0}|$ is the size of A_{s,j_0}.

The above conditions are mild and justifiable. Condition 1 is a commonly used condition that guarantees asymptotic normality of the ordinary maximum likelihood estimates. See, for example, Lehmann (1983). Condition 2(i) imposes some moment conditions on the latent factor and random errors and requires loading vectors are uniformly bounded away from infinity. In Condition 2(ii)–(iv), we impose some conditions on the covariates and the generalised loading matrix $\mathbf{D}^0$. They are all mild and normally required to establish the uniform consistency of the maximum likelihood estimates. Unlike some existing literature, for example, Bai and Li (2014), we do not impose any conditions on the estimators of variances

of random errors and latent factors. Hence, the conditions in this paper are more realistic in practice. Condition 3 imposes some restrictions on the homogeneity conditions. Condition 3(i) requires that the number of groups is fixed and the size of each group is on the same order. The scenario when $\mathcal{N}$ is diverging is not of interest in this paper as, even with a correct detection of homogeneity, the final estimation is still a diverging dimensional problem. In Condition 3(ii), we impose a lower bound on the smallest gap between two consecutive groups. Notice this lower bound can be any small positive constant and hence Condition 3(ii) will not cause any trouble in real application. Condition 3(iii) is a technical condition to avoid "tie" situations in change point detection and mainly used to reduce the ambiguousness in the theoretical proof. Condition 3(iv) requires $(n\log n)^2/T \to 0$ as $T \to \infty$. Condition 3(v) is a condition on the threshold δ. Condition 4 is a reasonable condition for distinguishing the "mis-specified" model from the true one. In general, all these technical conditions should be easily satisfied in most real applications.

8.2. *Proofs of the main theoretical results.* In this subsection, we give the detailed proofs of Theorems 1–3 introduced in Section 3.

PROOF OF THEOREM 1. To prove Theorem 1 is enough to show for any given constant vector $\mathbf{a} \in \mathbf{R}^p$ that

$$\sqrt{T}\big(\mathbf{a}^{\mathrm{T}}\boldsymbol{\Sigma}_i^*\mathbf{a}\big)^{-1/2}\mathbf{a}^{\mathrm{T}}\big(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\big) \xrightarrow{\mathrm{D}} N(0,1), \qquad i = 1,\dots,n, \tag{8.1}$$

where

$$\boldsymbol{\Sigma}_i^* = \lim_{T\to\infty} \Omega_i \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\big(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\big)^{-1}\mathcal{X}_t\right)^{-1} \Omega_i^{\mathrm{T}}. \tag{8.2}$$

To begin with, we introduce some notation used in this proof. For $i = 1,\dots,n$ and $t = 1,\dots,T$, we denote

$$\Omega_i = \big(\mathbf{0}_{p\times\{i(2p+1)-2p\}}\ \mathbf{I}_p\ \mathbf{0}_{p\times\{(n-i)(2p+1)+p\}}\big),$$

$$\mathbf{K} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\big(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}}\big)^{-1}\mathcal{X}_t \quad \text{and}$$

$$A_t = \mathcal{X}_t^{\mathrm{T}}\big(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}}\big)^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t = (A_{t,1},\dots,A_{t,n(2p+1)})^{\mathrm{T}}.$$

According to equation (8.48), for $i = 1,\dots,n$, we have

$$\big(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\big) = \Omega_i\mathbf{K}^{-1}\begin{pmatrix} \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} A_{t,1} \\ \vdots \\ \dfrac{1}{T}\displaystyle\sum_{t=1}^{T} A_{t,n(2p+1)} \end{pmatrix} \equiv \Omega_i\mathbf{K}^{-1}\begin{pmatrix} S_1 \\ \vdots \\ S_{n(2p+1)} \end{pmatrix}.$$

To make the structure of this proof clear, we divide the proof into three steps.

In the first step, we want to show

$$\sqrt{T}S_j \xrightarrow{\mathrm{D}} N(0,\sigma_j^{*2}), \qquad j=1,\ldots,n(2p+1), \tag{8.3}$$

and $\max_j \sigma_j^{*2} \le C$ for some sufficiently large positive constant C.

For $t=1,\ldots,T$, we can rewrite the vector A_t as

$$\begin{aligned} A_t &= \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t + \mathcal{X}_t^{\mathrm{T}}[(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)-\mathbf{I}](\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t \\ &\equiv \mathbf{U}_t + \mathbf{V}_t. \end{aligned}$$

We use $U_{t,j}$, $V_{t,j}$ to denote the jth entry of $\mathbf{U}_t$ and $\mathbf{V}_t$, respectively. Then we got

$$S_j = \frac{1}{T}\sum_{t=1}^{T} U_{t,j} + \frac{1}{T}\sum_{t=1}^{T} V_{t,j}, \qquad j=1,\ldots,n(2p+1). \tag{8.4}$$

Now we consider the first part of S_j. According to the model assumption, for each j, $U_{1,j},\ldots,U_{T,j}$ are i.i.d. random variables. By the strong law of large numbers, we have

$$\Pr\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}U_{t,j} = \mathrm{E}[U_{t,j}]\right) = 1.$$

So when $T\to\infty$, with probability equals to 1, we have

$$\max_j\big(\mathrm{E}[U_{t,j}]\big)^2 = \max_j\left[\frac{1}{T}\sum_{t=1}^{T}U_{t,j}\right]^2.$$

Similar to the way we proof Lemma 3 and notice the largest eigenvalue of $(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}$ is bounded by a positive constant C, we have

$$\max_j\left[\frac{1}{T}\sum_{t=1}^{T}U_{t,j}\right]^2 \le \left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{U}_t\right\|^2 \le C\left\|\frac{1}{T}\sum_{t=1}^{T}\mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\right\|^2 = O_p\left(\frac{n}{T}\right).$$

Therefore, we can show $\mathrm{E}[U_{t,j}]=0$ for $j=1,\ldots,n(2p+1)$ as

$$0 \le \big(\mathrm{E}[U_{t,j}]\big)^2 \le \max_j\big(\mathrm{E}[U_{t,j}]\big)^2 = \lim_{T\to\infty} O_p\left(\frac{n}{T}\right) = 0. \tag{8.5}$$

As $\mathrm{Var}(U_{t,j}) = E[(U_{t,j})^2] - (E[U_{t,j}])^2 = E[(U_{t,j})^2]$, again with the strong law of large numbers we have

$$\Pr\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}U_{t,j}^2 = \mathrm{E}\big[(U_{t,j})^2\big]\right) = 1.$$

Let $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^{\mathrm{T}}$ be the unit vector of length $n(2p+1)$ whose jth entry equals 1. Then for each t and j, we can show

$$\begin{aligned} U_{t,j}^2 &= \mathbf{e}_j^{\mathrm{T}}\mathbf{U}_t\mathbf{U}_t^{\mathrm{T}}\mathbf{e}_j = \mathbf{e}_j^{\mathrm{T}}\mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{\mathrm{T}}\mathbf{D}^0(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t\mathbf{e}_j \\ &\le C^2\mathbf{e}_j^{\mathrm{T}}(\mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t)(\mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t)^{\mathrm{T}}\mathbf{e}_j \\ &\le C^2\Big\{\max_i(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})^2 + \max_i\|(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})X_{it}\|^2 + \max_i\|\Gamma_i\mathbf{f}_t + \boldsymbol{\epsilon}_{it}\|^2\Big\}. \end{aligned}$$

The last inequality is due to $\mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t = (\mathbf{d}_{t1}^{\mathrm{T}}, \ldots, \mathbf{d}_{tn}^{\mathrm{T}})^{\mathrm{T}}$, where each $\mathbf{d}_{ti}$ is a $2p+1$ by 1 vector as introduced in (8.52).

Hence, when $T \to \infty$, with probability equals 1, $\mathrm{Var}(U_{t,j})$ can be uniformly upper bounded by a large enough positive constant as follows:

$$\begin{aligned} \mathrm{Var}(U_{t,j}) &\le C^2\Bigg\{\max_i\frac{1}{T}\sum_{t=1}^{T}(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})^2 + \frac{1}{T}\sum_{t=1}^{T}\max_i\|(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})X_{it}\|^2 \\ &\qquad + \frac{1}{T}\sum_{t=1}^{T}\max_i\|\Gamma_i\mathbf{f}_t + \boldsymbol{\epsilon}_{it}\|^2\Bigg\} \\ &\le C^2\Big\{\max_i(\|\boldsymbol{\lambda}_i\|^2 + \sigma_{1i}^2)\Big(1 + \max_{i,t}\|X_{it}\|^2\Big) + \max_i\|\Gamma_i\|_F^2 + p\max_i\sigma_{2i}^2\Big\} \\ &\le C^3 < \infty. \end{aligned} \tag{8.6}$$

With the results in (8.5) and (8.6), we showed when $T \to \infty$, with probability equals 1, for each $j \in \{1, \ldots, n(2p+1)\}$, $U_{1,j}, \ldots, U_{T,j}$ are i.i.d. random variables with $\mathrm{E}[U_{t,j}] = 0$, $\mathrm{Var}(U_{t,j}) \equiv \sigma_j^{*2}$ and $\max_j \sigma_j^{*2} \le C < \infty$ for some large enough positive constant C. Hence, by the central limit theorem, we have

$$\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}U_{t,j}\right) \xrightarrow{\mathrm{D}} N(0, \sigma_j^{*2}) \qquad \text{for } j = 1, \ldots, n(2p+1). \tag{8.7}$$

Then we consider the second part of S_j. According to Lemma 2(iii) and $n^2/T \to 0$, we have $\|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0) - \mathbf{I}\| = O_P(n/T) = o_P(1/\sqrt{T})$. We can show

$$\frac{1}{T}\sum_{t=1}^{T}V_{t,j} = \left(\frac{1}{T}\sum_{t=1}^{T}U_{t,j}\right)\cdot o_P(1/\sqrt{T}). \tag{8.8}$$

Combine the results of (8.4) and (8.8), we have for $j = 1, \ldots, n(2p+1)$

$$\sqrt{T}S_j = \left\{\sqrt{T}\left(\frac{1}{T}\sum_{t=1}^{T}U_{t,j}\right)\right\}\{1 + o_P(1)\}. \tag{8.9}$$

Then when $T \to \infty$, according (8.7), (8.9) and Slutsky's theorem, we finish the first step by showing the results we want in (8.3).

In the second step, we want to show the results in (8.1). Similar to the way we decompose A_t above, we can rewrite $\mathbf{K}$ as the sum of two parts

$$\mathbf{K} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t + \frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}[(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0) - \mathbf{I}](\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t$$

$$= \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t\right) \cdot \left\{1 + o_P\left(\frac{1}{\sqrt{T}}\right)\right\}.$$

According to Condition 2(iv), we denote $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t = \mathbf{K}^*$. Then we have $\mathbf{K} \longrightarrow^{\mathrm{P}} \mathbf{K}^*$ and $\mathbf{K}^{-1} \longrightarrow^{\mathrm{P}} \mathbf{K}^{*-1}$.

Denote $\mathbf{K}^{-1}$ and $\mathbf{K}^{*-1}$ by their row vectors: $\mathbf{K}^{-1} = (\eta_1, \ldots, \eta_{n(2p+1)})^{\mathrm{T}}$ and $\mathbf{K}^{*-1} = (\eta_1^*, \ldots, \eta_{n(2p+1)}^*)^{\mathrm{T}}$, where, for $k = 1, \ldots, n(2p+1)$, $\eta_k = (\eta_{k,1}, \ldots, \eta_{k,n(2p+1)})^{\mathrm{T}}$ and $\eta_k^* = (\eta_{k,1}^*, \ldots, \eta_{k,n(2p+1)}^*)^{\mathrm{T}}$ are the kth row of $\mathbf{K}^{-1}$ and $\mathbf{K}^{*-1}$, respectively.

Consider any constant vector $\mathbf{a} = (a_1, \ldots, a_p)^{\mathrm{T}} \in \mathbf{R}^p$, for $i = 1, \ldots, n$,

$$\begin{aligned}\mathbf{a}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0) &= \mathbf{a}^{\mathrm{T}}\Omega_i\mathbf{K}^{-1}(S_1, \ldots, S_{n(2p+1)})^{\mathrm{T}} \\ &= \mathbf{a}^{\mathrm{T}}(\eta_{i(2p+1)-2p+1}, \ldots, \eta_{i(2p+1)-p})(S_1, \ldots, S_{n(2p+1)})^{\mathrm{T}} \\ &= \sum_{j=1}^{p}\left\{a_j \sum_{k=1}^{n(2p+1)} \eta_{[i(2p+1)-2p+j],k} S_k\right\}.\end{aligned}$$

For each given i, j and k, as $\eta_{[i(2p+1)-2p+j],k} \longrightarrow^{\mathrm{P}} \eta^*_{[i(2p+1)-2p+j],k}$ and $\sqrt{T}S_j \longrightarrow^{\mathrm{D}} N(0, \sigma_j^{*2})$, by Slutsky's theorem we have

$$\sqrt{T}a_j\eta_{[i(2p+1)-2p+j],k}S_j \xrightarrow{\mathrm{D}} N\big(0, [a_j\eta^*_{[i(2p+1)-2p+j],k}\sigma_j^*]^2\big).$$

Hence $\sqrt{T}\mathbf{a}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0)$ can be considered as a sum of independent normal random variables. Suppose for any given finite n, we denote $\sqrt{T}\mathbf{a}^{\mathrm{T}}(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0) \equiv M_n$, it is easy to see $M_n \sim N(0, \sigma_n^{\dagger 2})$, where $\sigma_n^{\dagger 2} = \sum_{j=1}^{p}\sum_{k=1}^{n(2p+1)}[a_j \times \eta^*_{[i(2p+1)-2p+j],k}\sigma_j^*]^2$.

Furthermore if we define $\sigma^{\dagger 2} = \lim_{n\to\infty}\sigma_n^{\dagger 2}$, and notice the largest eigenvalue of $\mathbf{K}^{*-1}$ is upper bounded by a positive constant, we can show $\sigma^{\dagger 2}$ does exist and is bounded

$$\sigma^{\dagger 2} = \lim_{n\to\infty}\sigma_n^{\dagger 2} \le \max_j a_j^2 \cdot p\big\|\mathbf{K}^{*-1}\big\|^2 \cdot \max_j \sigma_j^{*2} \le C < \infty,$$

where C is a large enough positive constant.

Suppose we have $M \sim N(0, \sigma^{\dagger 2})$, and denote the characteristic functions of M_n and M by $\varphi_n(s)$ and $\varphi(s)$, respectively, it is easy to see

$$\begin{aligned}&\varphi_n(s) = \mathrm{e}^{-(1/2)\sigma_n^{\dagger 2}s^2}, \qquad \varphi(s) = \mathrm{e}^{-(1/2)\sigma^{\dagger 2}s^2} \quad \text{and} \\ &\varphi_n(s) \to \varphi(s) \qquad \text{as } \sigma_n^{\dagger 2} \to \sigma^{\dagger 2}.\end{aligned}$$

Then by Lévy's convergence theorem, we have

$$\sqrt{T}\mathbf{a}^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0) \xrightarrow{\mathrm{D}} N(0, \sigma^{\dagger 2}).$$

In the third step, we calculate $\sigma^{\dagger 2}$.

Notice $\mathrm{E}[\sqrt{T}\mathbf{a}^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0)] = 0$, when $n^2/T \to 0$, we have

$$\begin{aligned}\sigma^{\dagger 2} &= T \cdot \mathrm{E}\big[\mathbf{a}^{\mathrm{T}}\Omega_i\mathbf{K}^{*-1}[S_1, \ldots, S_{n(2p+1)}]^{\mathrm{T}}[S_1, \ldots, S_{n(2p+1)}]\mathbf{K}^{*-1}\Omega_i^{\mathrm{T}}\mathbf{a}\big] \\ &= T\mathbf{a}^{\mathrm{T}}\Omega_i\mathbf{K}^{*-1}\mathrm{E}\left[\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{U}_t\right]\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{U}_t\right]^{\mathrm{T}}\right]\mathbf{K}^{*-1}\Omega_i^{\mathrm{T}}\mathbf{a} \\ &\quad + 2\mathbf{a}^{\mathrm{T}}\Omega_i\mathbf{K}^{*-1}\mathrm{E}\left[\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\mathbf{U}_t\right]\mathrm{E}\left[\sqrt{T}\frac{1}{T}\sum_{t=1}^{T}\mathbf{U}_t\right]^{\mathrm{T}}\mathbf{K}^{*-1}\Omega_i^{\mathrm{T}}\mathbf{a} \\ &\equiv I_3 + I_4.\end{aligned}$$

Recall the definition of $\mathbf{U}_t$, and $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{\mathrm{T}} = \mathbf{I}$, when $n^2/T \to 0$, I_3 can be calculated as follows:

$$\begin{aligned}I_3 &= \mathbf{a}^{\mathrm{T}}\Omega_i\mathbf{K}^{*-1}\lim_{T\to\infty}\left[\frac{1}{T}\sum_{t=1}^{T}\mathcal{X}_t^{\mathrm{T}}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{\mathrm{T}}\mathbf{D}^0(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0)^{-1}\mathcal{X}_t\right]\mathbf{K}^{*-1}\Omega_i^{\mathrm{T}}\mathbf{a} \\ &= \mathbf{a}^{\mathrm{T}}\Omega_i\mathbf{K}^{*-1}\Omega_i^{\mathrm{T}}\mathbf{a}.\end{aligned}$$

On the other hand, using the results we obtained in (8.7), it is easy to see $I_4 = 0$ when $T \to 0$.

Combining the results form all three steps above, we complete the proof. □

PROOF OF THEOREM 2. Let $\tilde{\boldsymbol{\beta}}_i = (\tilde{\beta}_{i1}, \ldots, \tilde{\beta}_{ip})^{\mathrm{T}}$ be the initial estimator of $\boldsymbol{\beta}_i$ obtained in the initial estimation step. According to Lemma 3, we have $\tilde{\beta}_{ij} = \beta_{ij}^0 + O_p(\frac{1}{\sqrt{T}})$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Hence, we can write $\tilde{\beta}_{ij}$ as

$$\tilde{\beta}_{ij} = \beta_{ij}^0 + e_{ij}, \qquad e_{ij} = O_p\left(\frac{1}{\sqrt{T}}\right), \qquad i = 1, \ldots, n, j = 1, \ldots, p.$$

Then, one can expect the following holds in the probability sense:

$$\tilde{\beta}_{(s)} = \beta_{(s)}^0 + e_{(s)}, \qquad s = 1, \ldots, np,$$

where $e_{(s)}$'s can be viewed as the results of sorting e_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, from small to large. Note that

$$\max_{1\le t\le np}|e_{(s)}| = O_p\left(\frac{\log(np)}{\sqrt{T}}\right). \tag{8.10}$$

To make the presentation simple, in this proof, we assume there are two change points in the sequence $\{\beta_{(s)}^0, s = 1, \ldots, np\}$. The proof for the case with more than

two change points can be developed similarly. Hence, we consider the following homogeneity condition:

$$\beta_{(s)}^0 = \begin{cases} \mu_1, & \text{when } 1 \le s \le k_1^0, \\ \mu_2, & \text{when } k_1^0 + 1 \le s \le k_2^0, \\ \mu_3, & \text{when } k_2^0 + 1 \le s \le np. \end{cases} \tag{8.11}$$

Also, according to Condition 3(iii) and without loss of generality, we assume

$$S_{1,np}(k_1^0) < S_{1,np}(k_2^0). \tag{8.12}$$

Define

$$\hat{k}_1 = \underset{1 \le \kappa < np}{\arg\min}\, S_{1,np}(\kappa),$$

where

$$S_{1,np}(\kappa) = \frac{1}{np-1}\left\{ \sum_{s=1}^{\kappa} (\tilde{\beta}_{(s)} - \bar{\beta}_{1,\kappa})^2 + \sum_{t=\kappa+1}^{np} (\tilde{\beta}_{(s)} - \bar{\beta}_{\kappa+1,np})^2 \right\}$$

and

$$\bar{\beta}_{1,\kappa} = \frac{1}{\kappa}\sum_{t=1}^{\kappa} \tilde{\beta}_{(s)}, \qquad \bar{\beta}_{\kappa+1,np} = \frac{1}{np-\kappa}\sum_{t=\kappa+1}^{np} \tilde{\beta}_{(s)}.$$

As we can see, in this paper, the detection of homogeneity is equivalent to the detection of change points among $\tilde{\beta}_{(s)}, s = 1, \ldots, np$.

First, we want to show $P(\hat{k}_1 = k_1^0) \to 1$. To prove this, we only need to show $P(\hat{k}_1 < k_1^0) \to 0$, $P(k_1^0 < \hat{k}_1 \le k_2^0) \to 0$, and $P(k_2^0 < \hat{k}_1 \le np) \to 0$. In the following part, we will discuss these three scenarios one by one.

Scenario 1: $P(\hat{k}_1 < k_1^0) \to 0$. When $\kappa < k_1^0$, it is easy to see that

$$\bar{\beta}_{1,\kappa} = \mu_1 + \frac{1}{\kappa}\sum_{s=1}^{\kappa} e_{(s)} \quad \text{and}$$

$$\begin{aligned} \bar{\beta}_{\kappa+1,np} &= \frac{1}{np-\kappa}\left(\sum_{s=\kappa+1}^{k_1^0} \tilde{\beta}_{(s)} + \sum_{s=k_1^0+1}^{k_2^0} \tilde{\beta}_{(s)} + \sum_{s=k_2^0+1}^{np} \tilde{\beta}_{(s)} \right) \\ &= \frac{k_1^0 - \kappa}{np-\kappa}\mu_1 + \frac{k_2^0 - k_1^0}{np-\kappa}\mu_2 + \frac{np - k_2^0}{np-\kappa}\mu_3 + \frac{1}{np-\kappa}\sum_{s=\kappa+1}^{np} e_{(s)}. \end{aligned}$$

Thus,

$$\sum_{s=1}^{\kappa} (\tilde{\beta}_{(s)} - \bar{\beta}_{1,\kappa})^2 = \sum_{s=1}^{\kappa}\left(e_{(s)} - \frac{1}{\kappa}\sum_{s=1}^{\kappa} e_{(s)} \right)^2 = \sum_{s=1}^{\kappa} e_{(s)}^2 - \frac{1}{\kappa}\left(\sum_{s=1}^{\kappa} e_{(s)} \right)^2.$$

Denote

$$\tilde{\beta}_{(s)} - \bar{\beta}_{k+1,np} = \begin{cases} a_{np,\kappa} + e_{(s)} - \dfrac{1}{np-\kappa} \displaystyle\sum_{l=\kappa+1}^{np} e_{(l)}, & \text{if } s \in [1, k_1^0], \\ b_{np,\kappa} + e_{(s)} - \dfrac{1}{np-\kappa} \displaystyle\sum_{l=\kappa+1}^{np} e_{(l)}, & \text{if } s \in [\kappa+1, k_2^0], \\ c_{np,\kappa} + e_{(s)} - \dfrac{1}{np-\kappa} \displaystyle\sum_{l=\kappa+1}^{np} e_{(l)}, & \text{if } s \in [k_2^0+1, np], \end{cases}$$

where

$$a_{np,\kappa} = \frac{1}{np-\kappa}\big[(np-k_1^0)(\mu_1-\mu_2) + (np-k_2^0)(\mu_2-\mu_3)\big],$$

$$b_{np,\kappa} = \frac{1}{np-\kappa}\big[(k_1^0-\kappa)(\mu_2-\mu_1) + (np-k_2^0)(\mu_2-\mu_3)\big],$$

$$c_{np,\kappa} = \frac{1}{np-\kappa}\big[(k_1^0-\kappa)(\mu_2-\mu_1) + (k_2^0-\kappa)(\mu_3-\mu_2)\big].$$

We have

$$\begin{aligned} &\sum_{s=\kappa+1}^{np} (\tilde{\beta}_{(s)} - \bar{\beta}_{\kappa+1,np})^2 \\ &\quad = (k_1^0-\kappa)a_{np,\kappa}^2 + 2a_{np,\kappa} \sum_{s=\kappa+1}^{k_1^0} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right) \\ &\qquad + (k_2^0-k_1^0)b_{np,\kappa}^2 + 2b_{np,\kappa} \sum_{s=k_1^0+1}^{k_2^0} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right) \\ &\qquad + (np-k_2^0)c_{np,\kappa}^2 + 2c_{np,\kappa} \sum_{s=k_2^0+1}^{np} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right) \\ &\qquad + \sum_{s=\kappa+1}^{np} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right)^2. \end{aligned}$$

Hence, for $\kappa \le k_1^0$,

$$\begin{aligned} S_{1,np}(\kappa) &= \frac{k_1^0-\kappa}{np-1}a_{np,\kappa}^2 + \frac{k_2^0-k_1^0}{np-1}b_{np,\kappa}^2 + \frac{np-k_2^0}{np-1}c_{np,\kappa}^2 \\ &\quad + \frac{1}{np-1}\sum_{s=1}^{np} e_{(s)}^2 + R_1(\kappa), \end{aligned} \tag{8.13}$$

where

$$R_1(\kappa) = \frac{1}{np-1}\Bigg[2a_{np,\kappa}\sum_{s=\kappa+1}^{k_1^0} e_{(s)} + 2b_{np,\kappa}\sum_{s=k_1^0+1}^{k_2^0} e_{(s)} + 2c_{np,\kappa}\sum_{s=k_2^0+1}^{np} e_{(s)}\Bigg]$$
$$- \frac{2}{(np-1)(np-\kappa)}$$
$$\times \big[(k_1^0-\kappa)a_{np,\kappa} + (k_2^0-k_1)b_{np,\kappa} + (np-k_2^0)c_{np,\kappa}\big]\sum_{s=\kappa+1}^{np} e_{(s)}$$
$$- \frac{1}{(np-1)\kappa}\Bigg(\sum_{s=1}^{\kappa} e_{(s)}\Bigg)^2 - \frac{1}{(np-1)(np-\kappa)}\Bigg(\sum_{s=\kappa+1}^{np} e_{(s)}\Bigg)^2.$$

Note that $a_{np,\kappa}, b_{np,\kappa}$ and $c_{np,\kappa}$ are all bounded, recalling (8.10), one has

$$|R_1(\kappa)| = O_p\Big(\frac{\log(np)}{\sqrt{T}}\Big) \qquad \text{uniformly in } \kappa \in [1, k_1^0]. \tag{8.14}$$

Therefore, based on (8.13), (8.14) and some calculations, we can show

$$\begin{aligned} &S_{1,np}(\kappa) - S_{1,np}(k_1^0) \\ &\quad= \frac{1}{np-1}\cdot\frac{k_1^0-\kappa}{(1-\kappa/(np))(1-k_1^0/(np))} \\ &\qquad\times\Bigg[\Big(1-\frac{k_1^0}{np}\Big)(\mu_1-\mu_2) + \Big(1-\frac{k_2^0}{np}\Big)(\mu_2-\mu_3)\Bigg]^2 + O_p\Big(\frac{\log(np)}{\sqrt{T}}\Big) \\ &\quad=: \Delta_1(\kappa) + O_p\Big(\frac{\log(np)}{\sqrt{T}}\Big). \end{aligned} \tag{8.15}$$

Note that $\Delta_1(\kappa) > 0$ for any $1 \le \kappa < k_1^0$. Recall that p is fixed and $n\log n = o(\sqrt{T})$, therefore,

$$\frac{\log(np)}{\sqrt{T}} = o\Big(\frac{1}{np}\Big), \tag{8.16}$$

which implies that

$$\frac{\log(np)}{\sqrt{T}} = o_p\big(\Delta_1(\kappa)\big) \qquad \text{uniformly for } \kappa \in [1, k_1^0 - 1].$$

Based on the above arguments, one immediately has

$$\begin{aligned} P\big(\hat{k}_1 < k_1^0\big) &\le P\big(S_{1,np}(\hat{k}_1) - S_{1,np}(k_1^0) < 0, \hat{k}_1 < k_1^0\big) \\ &= P\Big(\Delta_1(\hat{k}_1) + O_p\Big(\frac{1}{\sqrt{T}}\Big) < 0, \hat{k}_1 < k_1^0\Big) \\ &\to 0. \end{aligned} \tag{8.17}$$

Scenario 2: $P(k_1^0 < \hat{k}_1 \le k_2^0) \to 0$. When $k_1^0 < \kappa \le k_2^0$, it is easy to see that

$$\bar{\beta}_{1,\kappa} = \frac{k_1^0}{\kappa}\mu_1 + \frac{\kappa - k_1^0}{\kappa}\mu_2 + \frac{1}{\kappa}\sum_{s=1}^{\kappa} e_{(s)} \quad \text{and}$$

$$\bar{\beta}_{\kappa+1,np} = \frac{k_2^0 - \kappa}{np - \kappa}\mu_2 + \frac{np - k_2^0}{np - \kappa}\mu_3 + \frac{1}{np - \kappa}\sum_{s=\kappa+1}^{np} e_{(s)}.$$

Thus,

$$\tilde{\beta}_{(s)} - \bar{\beta}_{1,\kappa} = \begin{cases} \dfrac{\kappa - k_1^0}{\kappa}(\mu_1 - \mu_2) + e_{(s)} - \dfrac{1}{\kappa}\displaystyle\sum_{l=1}^{\kappa} e_{(l)}, & \text{if } s \in [1, k_1^0], \\ \dfrac{k_1^0}{\kappa}(\mu_2 - \mu_1) + e_{(s)} - \dfrac{1}{\kappa}\displaystyle\sum_{l=1}^{\kappa} e_{(l)}, & \text{if } s \in [k_1^0 + 1, \kappa], \end{cases}$$

and

$$\tilde{\beta}_{(s)} - \bar{\beta}_{\kappa+1,np} = \begin{cases} \dfrac{np - k_2^0}{np - \kappa}(\mu_2 - \mu_3) + e_{(s)} - \dfrac{1}{np - \kappa}\displaystyle\sum_{l=\kappa+1}^{np} e_{(l)}, \\ \qquad \text{if } s \in [\kappa + 1, k_2^0], \\ \dfrac{k_2^0 - \kappa}{np - \kappa}(\mu_3 - \mu_2) + e_{(s)} - \dfrac{1}{np - \kappa}\displaystyle\sum_{l=\kappa+1}^{np} e_{(l)}, \\ \qquad \text{if } s \in [k_2^0 + 1, np]. \end{cases}$$

Therefore, for $k_1 < \kappa \le k_2$, one has

$$\begin{aligned} &\sum_{s=1}^{\kappa}(\tilde{\beta}_{(s)} - \bar{\beta}_{1,\kappa})^2 \\ &\quad = k_1^0 d_{np,\kappa}^2 + 2d_{np,\kappa}\sum_{s=1}^{k_1^0}\left(e_{(s)} - \frac{1}{\kappa}\sum_{l=1}^{\kappa} e_{(l)}\right) + \sum_{s=1}^{k_1^0}\left(e_{(s)} - \frac{1}{\kappa}\sum_{l=1}^{\kappa} e_{(l)}\right)^2 \\ &\qquad + (\kappa - k_1^0)e_{np,\kappa}^2 + 2e_{np,\kappa}\sum_{s=k_1^0+1}^{\kappa}\left(e_{(s)} - \frac{1}{\kappa}\sum_{l=1}^{\kappa} e_{(l)}\right) \\ &\qquad + \sum_{s=k_1^0+1}^{\kappa}\left(e_{(s)} - \frac{1}{\kappa}\sum_{l=1}^{\kappa} e_{(l)}\right)^2, \end{aligned}$$

where

$$d_{np,\kappa} = \frac{\kappa - k_1^0}{\kappa}(\mu_1 - \mu_2), \qquad e_{np,\kappa} = \frac{k_1^0}{\kappa}(\mu_2 - \mu_1).$$

Also one has

$$\begin{aligned}
&\sum_{s=\kappa+1}^{np} (\tilde{\beta}_{(s)} - \bar{\beta}_{\kappa+1,np})^2 \\
&\quad = (k_2^0 - \kappa) f_{np,\kappa}^2 + 2 f_{np,\kappa} \sum_{s=\kappa+1}^{k_2^0} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right) \\
&\qquad + \sum_{s=\kappa+1}^{k_2^0} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right)^2 + (np - k_2^0) g_{np,\kappa}^2 \\
&\qquad + 2 g_{np,\kappa} \sum_{s=k_2^0+1}^{np} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right) \\
&\qquad + \sum_{s=k_2^0+1}^{np} \left(e_{(s)} - \frac{1}{np-\kappa} \sum_{l=\kappa+1}^{np} e_{(l)} \right)^2,
\end{aligned}$$

where

$$f_{np,\kappa} = \frac{np - k_2^0}{np - \kappa} (\mu_2 - \mu_3), \qquad g_{np,\kappa} = \frac{k_2^0 - \kappa}{np - \kappa} (\mu_3 - \mu_2).$$

Thus,

$$\begin{aligned}
S_{1,np}(\kappa) &= \frac{k_1^0}{np-1} d_{np,\kappa}^2 + \frac{\kappa - k_1^0}{np-1} e_{np,\kappa}^2 + \frac{k_2^0 - \kappa}{np-1} f_{np,\kappa}^2 + \frac{np - k_2^0}{np-1} g_{np,\kappa}^2 \\
&\quad + \frac{1}{np-1} \sum_{s=1}^{np} e_{(s)}^2 + R_2(\kappa) \\
&= \frac{k_1^0 (\kappa - k_1^0)}{\kappa (np-1)} (\mu_2 - \mu_1)^2 + \frac{(k_2^0 - \kappa)(np - k_2^0)}{(np-1)(np-\kappa)} (\mu_3 - \mu_2)^2 \\
&\quad + \frac{1}{np-1} \sum_{s=1}^{np} e_{(s)}^2 + R_2(\kappa),
\end{aligned} \tag{8.18}$$

where

$$\begin{aligned}
R_2(\kappa) = \frac{1}{np-1} \cdot \Bigg[& 2 d_{np,\kappa} \sum_{s=1}^{k_1^0} e_{(s)} + 2 e_{np,\kappa} \sum_{s=k_1^0+1}^{\kappa} e_{(s)} + 2 f_{np,\kappa} \sum_{s=\kappa+1}^{k_2^0} e_{(s)} \\
& + 2 g_{np,\kappa} \sum_{s=k_2^0+1}^{np} e_{(s)} \Bigg] - \frac{2}{np-1} \left[k_1^0 d_{np,\kappa} + (\kappa - k_1^0) e_{np,\kappa} \right] \cdot \frac{1}{\kappa} \sum_{s=1}^{\kappa} e_{(s)}
\end{aligned} \tag{8.19}$$

$$- \frac{1}{\kappa(np-1)}\Bigg(\sum_{s=1}^{\kappa} e_{(s)}\Bigg)^2 - \frac{1}{(np-\kappa)(np-1)}\Bigg(\sum_{s=\kappa+1}^{np} e_{(s)}\Bigg)^2$$
$$- \frac{2}{np-1}\big[(k_2^0-\kappa) f_{np,\kappa} + (np-k_2^0) g_{np,\kappa}\big] \cdot \frac{1}{np-\kappa} \sum_{s=\kappa+1}^{np} e_{(s)}.$$

Note that $d_{np,\kappa}$, $e_{np,\kappa}$, $f_{np,\kappa}$ and $g_{np,\kappa}$ are all bounded, recalling (8.10), we have

$$|R_2(\kappa)| = O_p\Big(\frac{\log(np)}{\sqrt{T}}\Big) \qquad \text{uniformly in } \kappa \in [k_1+1, k_2]. \tag{8.20}$$

Therefore, based on (8.18), (8.20) and some calculations, we can show

$$\begin{aligned} &S_{1,np}(\kappa) - S_{1,np}(k_1^0) \\ &\quad = \frac{\kappa - k_1^0}{np-1}\Big[\frac{k_1^0}{\kappa}(\mu_2-\mu_1)^2 - \frac{(np-k_2^0)^2}{(np-\kappa)(np-k_1^0)}(\mu_3-\mu_2)^2\Big] \\ &\qquad + O_p\Big(\frac{\log(np)}{\sqrt{T}}\Big) \\ &\quad =: \Delta_2(\kappa) + O_p\Big(\frac{\log(np)}{\sqrt{T}}\Big). \end{aligned} \tag{8.21}$$

Note that $(np-k_2^0)/(np-\kappa) < k_2^0/\kappa$, we have

$$\begin{aligned} \Delta_2(\kappa) &= \frac{\kappa - k_1^0}{np-1}\Big[\frac{k_1^0}{\kappa}(\mu_2-\mu_1)^2 - \frac{(np-k_2^0)^2}{(np-\kappa)(np-k_1^0)}(\mu_3-\mu_2)^2\Big] \\ &\geq \frac{\kappa - k_1^0}{np-1} \cdot \frac{k_2^0}{\kappa} \cdot \Big[\frac{k_1^0}{k_2^0}(\mu_2-\mu_1)^2 - \frac{np-k_2^0}{np-k_1^0}(\mu_3-\mu_2)^2\Big]. \end{aligned} \tag{8.22}$$

According to (8.11), (8.13) and (8.18) and notice that $\frac{1}{np-1}\sum_{s=1}^{np} e_{(s)}^2 = o_p(1)$, we can see that (8.12) is equivalent to

$$\frac{k_1^0}{k_2^0}(\mu_2-\mu_1)^2 > \frac{1-k_2^0}{1-k_1^0}(\mu_3-\mu_2)^2.$$

Hence, $\Delta_2(\kappa) > 0$ for any $k_1^0 < \kappa \leq k_2^0$ when n and T is large. Moreover, it follows from (8.16) that

$$\frac{\log(np)}{\sqrt{T}} = o\big(\Delta_2(\kappa)\big) \qquad \text{uniformly in } k_1^0 < \kappa \leq k_2^0.$$

Base on the above arguments, we have

$$
\begin{aligned}
P\big(k_1^0 < \hat{k}_1 \le k_2^0\big) &\le P\big(S_{1,np}(\hat{k}_1) - S_{1,np}(k_1^0) < 0, k_1^0 < \hat{k}_1 \le k_2^0\big) \\
&= P\bigg(\Delta_2(\hat{k}_1) + O_p\bigg(\frac{\log(np)}{\sqrt{T}}\bigg) < 0, k_1^0 < \hat{k}_1 \le k_2^0\bigg) \\
&\to 0.
\end{aligned} \tag{8.23}
$$

Scenario 3: $P(k_2^0 < \hat{k}_1 \le np) \to 0$. When $k_2^0 < \kappa \le np$, symmetric with (8.13) in the Scenario 1, we have

$$
\begin{aligned}
S_{1,np}(\kappa) = {} & \frac{k_1^0}{np-1} h_{np,\kappa}^2 + \frac{k_2^0 - k_1^0}{np-1} p_{np,\kappa}^2 + \frac{\kappa - k_2^0}{np-1} q_{np,\kappa}^2 \\
& + \frac{1}{np-1} \sum_{s=1}^{np} e_{(s)}^2 + R_3(\kappa),
\end{aligned} \tag{8.24}
$$

where

$$
\begin{aligned}
h_{np,\kappa} &= \frac{1}{\kappa}\big[(\kappa - k_1^0)(\mu_1 - \mu_2) + (\kappa - k_2^0)(\mu_2 - \mu_3)\big], \\
p_{np,\kappa} &= \frac{1}{\kappa}\big[k_1^0(\mu_2 - \mu_1) + (\kappa - k_2^0)(\mu_2 - \mu_3)\big], \\
q_{np,\kappa} &= \frac{1}{\kappa}\big[k_1^0(\mu_2 - \mu_1) + k_2^0(\mu_3 - \mu_2)\big],
\end{aligned}
$$

and

$$
\begin{aligned}
R_3(\kappa) = {} & \frac{1}{np-1}\Bigg[2h_{np,\kappa} \sum_{s=1}^{k_1^0} e_{(s)} + 2p_{np,\kappa} \sum_{s=k_1^0+1}^{k_2^0} e_{(s)} + 2q_{np,\kappa} \sum_{s=k_2^0+1}^{\kappa} e_{(s)}\Bigg] \\
& - \frac{2}{np-1}\big[k_1^0 h_{np,\kappa} + (k_2^0 - k_1^0) p_{np,\kappa} + (np - k_2^0) q_{np,\kappa}\big] \sum_{s=1}^{\kappa} e_{(s)} \\
& - \frac{1}{(np-1)(np-\kappa)}\Bigg(\sum_{s=\kappa+1}^{np} e_{(s)}\Bigg)^2 - \frac{1}{(np-1)\kappa}\Bigg(\sum_{s=1}^{\kappa} e_{(s)}\Bigg)^2.
\end{aligned}
$$

First, taking $\kappa = k_2^0$ in (8.21) and (8.22) respectively, we have

$$
S_{1,np}(k_2^0) - S_{1,np}(k_1^0) = \Delta_2(k_2^0) + O_p\bigg(\frac{\log(np)}{\sqrt{T}}\bigg) \tag{8.25}
$$

and for large n and T,

$$
\Delta_2(k_2^0) \ge \frac{1}{2}(\tau_2^0 - \tau_1^0)\bigg[\frac{\tau_1^0}{\tau_2^0}(\mu_2 - \mu_1)^2 - \frac{1 - \tau_2^0}{1 - \tau_1^0}(\mu_3 - \mu_2)^2\bigg] =: C^* > 0. \tag{8.26}
$$

Second, notice that (8.15) implies that there exists a constant $M > 0$ such that

$$S_{1,np}(\kappa) - S_{1,np}(k_1^0) \geq -\frac{M\log(np)}{\sqrt{T}} \qquad \text{in probability when } \kappa \in [1, k_1^0],$$

which, by symmetry, implies that there exists a constant $M' > 0$ such that

$$S_{1,np}(\kappa) - S_{1,np}(k_2^0) \geq -\frac{M'\log(np)}{\sqrt{T}} \tag{8.27}$$

in probability when $\kappa \in [k_2^0 + 1, np]$.

It follows from (8.25), (8.26) and (8.27) that there exists a constant $C_0 > 0$ such that, when $\kappa \in [k_2^0 + 1, np]$ and n and T are large,

$$\begin{aligned} S_{1,np}(\kappa) - S_{1,np}(k_1^0) &= S_{1,np}(\kappa) - S_{1,np}(k_2^0) + S_{1,np}(k_2^0) - S_{1,np}(k_1^0) \\ &\geq C^* - \frac{C_0\log(np)}{\sqrt{T}} \qquad \text{in probability.} \end{aligned}$$

This inequality immediately yields

$$P(\hat{k}_1 > k_2^0) \leq P\bigl(S_{1,np}(\hat{k}_1) - S_{1,np}(k_1^0) < 0, \hat{k}_1 > k_2^0\bigr) \to 0. \tag{8.28}$$

Now, combining (8.17), (8.23) and (8.28) together, we have

$$P(\hat{k}_1 = k_1^0) \to 1. \tag{8.29}$$

Given this consistency result, we can sort and divide the sequence $\{\tilde{\beta}_{(1)}, \ldots, \tilde{\beta}_{(np)}\}$ into two subregions with the first subregion consisting of the first $\hat{k}_1$ sorted initial estimators and the second subregion consisting of the rest $np - \hat{k}_1$ sorted initial estimators. According to Condition 3, estimating the second change point based on the rest $np - \hat{k}_1$ sorted initial estimators is equivalent to estimating one change point based on $\{\tilde{\beta}_{(k_1^0+1)}, \ldots, \tilde{\beta}_{(np)}\}$. The consistency of the estimator of the second change point, that is, $\hat{k}_2$, can be proved by a similar fashion as (8.29).

Based on Condition 3 and the above arguments, one can estimate change points $k_{i_1}^0, \ldots, k_{i_m}^0$ consistently through an iterative algorithm.

The number of change points $\mathcal{N}$ can also be estimated consistently by our detection of homogeneity method. In fact, by Lemma 3, it is easy to see that

$$\min_{1\leq s\leq \hat{\mathcal{N}}} (\tilde{\beta}_{(\hat{k}(s+1))} - \tilde{\beta}_{(\hat{k}(s))})^2 = \min_{1\leq k\leq \mathcal{N}} (\mu_{k+1} - \mu_k)^2 + o_p(1),$$

where μ_k, $k = 1, \ldots, \mathcal{N} + 1$, is the sorted $\beta_{0,k}$, $k = 1, \ldots, \mathcal{N} + 1$.

If the subregion $\{\tilde{\beta}_{(i)}, \ldots, \tilde{\beta}_{(j)}\}$ contains at least one change point, it follows from the similar arguments in (8.24) that

$$S_{i,j}(j) \geq C\frac{j-i}{np-1} \min_{1\leq k\leq \mathcal{N}} (\mu_{k+1} - \mu_k)^2 + o_p(1) \tag{8.30}$$

as long as $j - i$ has the same order as np, where C is some positive constant. According to Condition 3(v), the right-hand side of the above inequality is on the constant order while the threshold $\delta \to 0$ as $np \to \infty$, hence we will successfully detect one true change point in this subregion with probability approaching one. Obviously, other true change points will also be successfully detected one by one by the iterative method proposed in step 2. If there is no change point left in the subregion $\{\tilde{\beta}_{(i)}, \ldots, \tilde{\beta}_{(j)}\}$, then it is clear that $S_{i,j}(j)$ will approach to zero with rate $o(1/(np))$ by (8.10). Note that from Condition 3(v), $np\delta \to \infty$ as $np \to \infty$. Hence, our procedure will stop the detection correctly in this subregion. The above arguments show that $\mathcal{N}$ will be estimated consistently by our procedure.

After re-parameterising β_{ij} based on the detection of homogeneity, we are able to show

$$P(\hat{\mathcal{N}} = \mathcal{N}) \longrightarrow 1 \quad \text{and} \quad P\big(\hat{k}_{(s)} = k^0_{(s)}\big) \longrightarrow 1, \qquad s = 1, \ldots, \mathcal{N}.$$

As the subspace $\mathcal{M}_A$ defined in Section 3 is homotopy equivalent to $\mathbb{R}^{\mathcal{N}+1}$, $\mathcal{M}_A$ is a connected space. Then similar to Lemma 1, we can show there exists a local maximiser $\widehat{\boldsymbol{\Theta}}^{\text{oracle}}$ of $L(\boldsymbol{\Theta})$ on $\mathcal{M}_A$ and $\|\widehat{\boldsymbol{\theta}}^{\text{oracle}} - \boldsymbol{\theta}_0\| = O_P(\sqrt{\frac{n}{T}})$. Also similar to Lemma 2, we have all the eigenvalues of $[(\widehat{\mathbf{D}}^{\text{oracle}})^{\mathrm{T}}\widehat{\mathbf{D}}^{\text{oracle}}]$ are bounded by constant and $\|[(\widehat{\mathbf{D}}^{\text{oracle}})^{\mathrm{T}}\widehat{\mathbf{D}}^{\text{oracle}}]^{-1}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0) - \mathbf{I}\|^2 = O_P(\frac{n}{T})$.

In the proof above, we showed with probability approaching one our proposed homogeneity detection method can correctly estimate the number of groups and the positions of all change points. Hence, we can claim with probability approaching one the estimator $\widehat{\boldsymbol{\beta}}$ obtained from the final estimation step equals the oracle estimator $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$.

Recall the definition of the subspace $\mathcal{M}_A$ and matrix Ψ_A, it is easy to see both $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ and $\boldsymbol{\beta}^0$ belong to $\mathcal{M}_A$, and we have the following relationships:

$$\boldsymbol{\beta}^0_A = \Psi_A \boldsymbol{\Omega} \boldsymbol{\theta}^0 \quad \text{and} \quad \widehat{\boldsymbol{\beta}}^{\text{oracle}}_A = \Psi_A \boldsymbol{\Omega} \widehat{\boldsymbol{\beta}}^{\text{oracle}}, \tag{8.31}$$

where $\boldsymbol{\Omega} = (\Omega_1, \ldots, \Omega_n)^{\mathrm{T}}$ and Ω_i, $i = 1, \ldots, n$ is defined in (8.47).

Using the relationships in (8.31) and similar to (8.48), we have

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{oracle}}_A - \boldsymbol{\beta}^0_A &= \Psi_A \boldsymbol{\Omega} \mathbf{K}^{\dagger -1}\left[\frac{1}{T}\sum_{t=1}^{T} A_t^{\dagger}\right] = \Psi_A \begin{pmatrix} \Omega_1 \mathbf{K}^{\dagger -1}\left[\frac{1}{T}\sum_{t=1}^{T} A_t^{\dagger}\right] \\ \vdots \\ \Omega_n \mathbf{K}^{\dagger -1}\left[\frac{1}{T}\sum_{t=1}^{T} A_t^{\dagger}\right] \end{pmatrix} \\ &\equiv \Psi_A \boldsymbol{\Xi},\end{aligned}$$

where $\mathbf{K}^{\dagger} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}[(\widehat{\mathbf{D}}^{\text{oracle}})^{\mathrm{T}}\widehat{\mathbf{D}}^{\text{oracle}}]^{-1}\mathcal{X}_t$, $A_t^{\dagger} = \mathcal{X}_t^{\mathrm{T}}[(\widehat{\mathbf{D}}^{\text{oracle}})^{\mathrm{T}}\widehat{\mathbf{D}}^{\text{oracle}}]^{-1} \times \mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t$ for $t = 1, \ldots, T$, $\boldsymbol{\Xi} = (\boldsymbol{\Xi}_1^{\mathrm{T}}, \ldots, \boldsymbol{\Xi}_n^{\mathrm{T}})^{\mathrm{T}}$, and $\boldsymbol{\Xi}_i = (\Xi_{i1}, \ldots, \Xi_{ip})^{\mathrm{T}}$ for $i = 1, \ldots, n$.

Follow the same way we prove Theorem 1, and we are able to show for $i = 1, \ldots, n$, $j = 1, \ldots, p$

$$\sqrt{T}\,\Xi_{ij} \xrightarrow{\mathrm{D}} N\bigl(0, \sigma_{ij}^{*2}\bigr), \tag{8.32}$$

where σ_{ij}^{*2} is the jth diagonal entry of

$$\boldsymbol{\Sigma}_i^* = \lim_{T\to\infty} \Omega_i \left(\frac{1}{T} \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}} \bigl(\mathbf{D}^{0\mathrm{T}} \mathbf{D}^0\bigr)^{-1} \mathcal{X}_t \right)^{-1} \Omega_i^{\mathrm{T}}. \tag{8.33}$$

Furthermore, all σ_{ij}^{*2}s are uniformly upper bounded by some large enough positive constant C, that is, $\max_{i,j} \sigma_{ij}^{*2} < C$.

Then for any given constant vector $\mathbf{a}^\dagger \in \mathbb{R}^{\mathcal{N}+1}$, when $T \to \infty$, $\sqrt{nT}\mathbf{a}^{\dagger\mathrm{T}} \times (\widehat{\boldsymbol{\beta}}_A^{\mathrm{oracle}} - \boldsymbol{\beta}_A^0)$ can be treated as a summation of np independent zero mean normal random variables. In addition, the limit of the summation of the variances of these normal random variables exists and is bounded. Hence, follow the similar way we prove the Theorem 1, we can complete the proof by constructing characteristic functions and using Lévy's convergence theorem. □

PROOF OF THEOREM 3. Consider any vector $\mathbf{a} \in \mathbb{R}^{np}$, we can number the indexes of its entries as $\mathbf{a} = \{\mathbf{a}_1^{\mathrm{T}}, \ldots, \mathbf{a}_n^{\mathrm{T}}\}^{\mathrm{T}}$, where $\mathbf{a}_i = \{a_{i1}, \ldots, a_{ip}\}^{\mathrm{T}}$ for $i = 1, \ldots, n$. Let $\mathcal{M}_p$ be a subspace of $\mathbb{R}^{np}$ defined by

$$\mathcal{M}_p = \bigl\{\mathbf{a} \in \mathbb{R}^{np} : a_{ij} = a_{kj}, \text{ for any } i, k = 1, \ldots, n \text{ and } j = 1, \ldots, p\bigr\}. \tag{8.34}$$

According to the definition of $\check{\boldsymbol{\beta}}$, it is easy to see $\check{\boldsymbol{\beta}} \in \mathcal{M}_p$.

Then we can define a map $\mathcal{T} : \mathbb{R}^{np} \to \mathcal{M}_p$, such that for any $\mathbf{a} = \{\mathbf{a}_1^{\mathrm{T}}, \ldots, \mathbf{a}_n^{\mathrm{T}}\}^{\mathrm{T}} \in \mathbb{R}^{np}$, $\mathcal{T}(\mathbf{a}) = (\mu_{\mathbf{a}}^{\mathrm{T}}, \ldots, \mu_{\mathbf{a}}^{\mathrm{T}})^{\mathrm{T}}$, where $\mu_{\mathbf{a}} = \frac{1}{n}\sum_{i=1}^n \mathbf{a}_i$. It is easy to check, $\mathcal{T}$ is the orthogonal projection from $\mathbb{R}^{np}$ to $\mathcal{M}_p$.

Suppose there exists a pair of index a, b such that $\beta_{aj}^0 \neq \beta_{bj}^0$ for some $a, b = 1, \ldots, n$, $a \neq b$ and $j = 1, \ldots, p$. Without loss of generality, here we assume $\beta_{aj}^0 < \beta_{bj}^0$. Then the relationship between β_{aj}^0, β_{bj}^0 and $\frac{1}{n}\sum_{i=1}^n \beta_{ij}^0$ is one of the following four scenarios:

$$\text{(i)}\quad \frac{1}{n}\sum_{i=1}^n \beta_{ij}^0 \le \beta_{aj}^0; \qquad \text{(ii)}\quad \beta_{aj}^0 \le \frac{1}{n}\sum_{i=1}^n \beta_{ij}^0 \le \beta_{aj}^0 + \frac{1}{2}\mathcal{J};$$

$$\text{(iii)}\quad \beta_{aj}^0 + \frac{1}{2}\mathcal{J} < \frac{1}{n}\sum_{i=1}^n \beta_{ij}^0 \le \beta_{bj}^0; \qquad \text{(iv)}\quad \beta_{bj}^0 < \frac{1}{n}\sum_{i=1}^n \beta_{ij}^0.$$

One can see, for all four scenarios above, we have

$$\max\left\{ \left| \beta_{aj}^0 - \frac{1}{n}\sum_{i=1}^n \beta_{ij}^0 \right|, \left| \beta_{bj}^0 - \frac{1}{n}\sum_{i=1}^n \beta_{ij}^0 \right| \right\} \ge \frac{1}{2}\mathcal{J}. \tag{8.35}$$

According to Condition 4, it is easy to see $\boldsymbol{\beta}^0 \notin \mathcal{M}_p$. Furthermore, there exists an integer j_0, $1 \le j_0 \le p$, such that there are at least $\min\{\mathcal{G}_n n, (1-\mathcal{G}_n)n\}$ pairs of elements in $\{\beta^0_{1,j_0}, \ldots, \beta^0_{n,j_0}\}$ that have different values.

Then using the property of orthogonal projection and (8.35), we can complete the proof by showing

$$\begin{aligned}
\|\check{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2 &\ge \|\boldsymbol{\beta}^0 - \mathcal{T}(\boldsymbol{\beta}^0)\|^2 = \sum_{j=1}^{p}\sum_{i=1}^{n}\left|\beta^0_{ij} - \frac{1}{n}\sum_{i=1}^{n}\beta^0_{ij}\right|^2 \\
&\ge \sum_{i=1}^{n}\left|\beta^0_{i,j_0} - \frac{1}{n}\sum_{i=1}^{n}\beta^0_{i,j_0}\right|^2 \\
&\ge \min\{\mathcal{G}_0 n, (1-\mathcal{G}_1)n\}\max\left\{\left|\beta^0_{aj} - \frac{1}{n}\sum_{i=1}^{n}\beta^0_{ij}\right|^2, \left|\beta^0_{bj} - \frac{1}{n}\sum_{i=1}^{n}\beta^0_{ij}\right|^2\right\} \\
&\ge \mathcal{C}^2 n \qquad \text{where } \mathcal{C}^2 = \frac{\mathcal{J}^2 \min\{\mathcal{G}_0, 1-\mathcal{G}_1\}}{4}.
\end{aligned}$$

□

8.3. *Proofs of some technical lemmas.* In this subsection, we introduce some useful technical lemmas and their proofs.

LEMMA 1. *Under Condition 1 in Section 8.1, there exists a local maximiser* $\tilde{\boldsymbol{\Theta}}$ *of* $L(\boldsymbol{\Theta})$, *and* $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_P(\sqrt{\frac{n}{T}})$.

PROOF. As $\boldsymbol{\theta}$ is a sub-vector of $\boldsymbol{\Theta}$, we have $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| \le \|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\|$. Let $r_{nT} = (n/T)^{-1/2}$. Lemma 1 is proven if one can show there exists a local maximiser of $L(\boldsymbol{\Theta})$ such that $\|\tilde{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_0\| = O_P(r_{nT})$. This implies, for any given $\epsilon > 0$, there exists a large positive constant C such that

$$P\Big\{\sup_{\|\mathbf{u}\|=C} L(\boldsymbol{\Theta}_0 + r_{nT}\mathbf{u}) < L(\boldsymbol{\Theta}_0)\Big\} \ge 1 - \epsilon. \tag{8.36}$$

Let $L'(\boldsymbol{\Theta}_0)$ be the gradient vector of $L(\boldsymbol{\Theta})$ at $\boldsymbol{\Theta} = \boldsymbol{\Theta}_0$, and $L'(\boldsymbol{\Theta}_0)_j$ be the jth component of $L'(\boldsymbol{\Theta}_0)$ for $j = 1, \ldots, d_n$ where $d_n = n\{(p+1)(q+2)+1\}$. By the standard argument on Taylor's expansion of the likelihood function, we have

$$\begin{aligned}
&L(\boldsymbol{\Theta}_0 + r_{nT}\mathbf{u}) - L(\boldsymbol{\Theta}_0) \\
&\quad = r_{nT} L'(\boldsymbol{\Theta}_0)^{\mathrm{T}}\mathbf{u} - \tfrac{1}{2} T r_{nT}^2 \mathbf{u}^{\mathrm{T}} I(\boldsymbol{\Theta}_0)\mathbf{u}\{1 + o_P(1)\}.
\end{aligned} \tag{8.37}$$

It is straightforward that the first term in (8.37) can be written as

$$r_{nT} L'(\boldsymbol{\Theta}_0)^{\mathrm{T}}\mathbf{u} = C r_{nT} \sum_{j=1}^{d_n} L'(\boldsymbol{\Theta}_0)_j \frac{u_j}{C} \equiv C r_{nT} M_n,$$

where u_j is the jth component of $\mathbf{u}$.

By the central limit theorem, we have $T^{-1/2}L'(\boldsymbol{\Theta}_0)_j \longrightarrow^{\mathrm{D}} N(0, \sigma_j^{*2})$ for $j = 1, \ldots, d_n$, and $\max_j \sigma_j^{*2}$ is upper bounded. Then we can see $T^{-1/2}M_n$ is a summation of independent normal random variables with zero mean and uniformly bounded variances. We define $\sigma_n^{\dagger 2} = \mathrm{Var}(T^{-1/2}n^{-1/2}M_n)$ and $\sigma^{\dagger 2} = \lim_{n\to\infty} \sigma_n^{\dagger 2}$. Notice $|u_j/C| \le 1$ for all j, we can show the limit $\sigma^{\dagger 2}$ exists and is upper bounded as follows:

$$\begin{aligned}
\sigma^{\dagger 2} &= \lim_{n\to\infty} n^{-1} \sum_{j=1}^{d_n} \mathrm{Var}\left[T^{-1/2} L'(\boldsymbol{\Theta}_0)_j \frac{u_j}{C} \right] \\
&= \lim_{n\to\infty} n^{-1} \sum_{j=1}^{d_n} \sigma_j^{*2} \left(\frac{u_j}{C} \right)^2 \\
&\le \lim_{n\to\infty} \{(p+1)(q+2)+1\} \max_j \sigma_j^{*2} \le C^*,
\end{aligned}$$

where C^* is a large enough positive constant.

Suppose we have $T^{-1/2}n^{-1/2}M \sim N(0, \sigma^{\dagger 2})$, and denote the characteristic functions of $T^{-1/2}n^{-1/2}M_n$ and $T^{-1/2}n^{-1/2}M$ by $\varphi_n(s)$ and $\varphi(s)$, respectively. It is easy to see

$$\begin{aligned}
&\varphi_n(s) = \mathrm{e}^{-(1/2)\sigma_n^{\dagger 2} s^2}, \qquad \varphi(s) = \mathrm{e}^{-(1/2)\sigma^{\dagger 2} s^2} \quad \text{and} \\
&\varphi_n(s) \to \varphi(s) \qquad \text{as } \sigma_n^{\dagger 2} \to \sigma^{\dagger 2}.
\end{aligned}$$

Then by Lévy's convergence theorem, we have

$$T^{-1/2}n^{-1/2}M_n \xrightarrow{\mathrm{D}} N\big(0, \sigma^{\dagger 2}\big).$$

Furthermore, notice $\sigma^{\dagger 2}$ is bounded, we have $T^{-1/2}n^{-1/2}M_n = O_P(1)$. So we can see, the first term in (8.37) is a random variable that on the order $O_p(r_{nT}n^{1/2}T^{1/2}) = O_P(n) = O_P(r_{nT}^2 T)$. Therefore, by choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. Thus, we complete the proof by showing (8.36) holds. □

In the following lemma, we use $\tilde{\lambda}_{\min}$, $\tilde{\lambda}_{\max}$ and $\lambda_{\min}^0$, $\lambda_{\max}^0$ to denote the smallest and largest eigenvalues of $\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}}$ and $\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0$, respectively. For any matrix A, we use A^+ to denote the Moore–Penrose pseudo-inverse of A.

LEMMA 2. *Under the conditions of Lemma 1 and Condition 2 in Section 8.1*:

(i) $\|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\|^2 = O_P(\frac{n}{T})$;
(ii) $C^{-1} \le \tilde{\lambda}_{\min} \le \tilde{\lambda}_{\max} \le C$, *where C is a large enough positive constant*;
(iii) $\|(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0) - \mathbf{I}\|^2 = O_P(\frac{n}{T})$.

PROOF. Given $\widetilde{\boldsymbol{\theta}}$, the first-order condition of $\mathbf{D}$ can be derived through

$$\frac{\partial L(\widetilde{\boldsymbol{\theta}}, \mathbf{D})}{\partial \mathbf{D}} = \frac{\partial H_1(\widetilde{\boldsymbol{\theta}}, \mathbf{D})}{\partial \mathbf{D}} + \frac{\partial H_2(\mathbf{D})}{\partial \mathbf{D}} = 0 \qquad \text{where}$$

$$H_1(\widetilde{\boldsymbol{\theta}}, \mathbf{D}) = -\frac{1}{2}\sum_{t=1}^{T} \operatorname{tr}\{(\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}})^{\mathrm{T}}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}(\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}})\} \quad \text{and} \tag{8.38}$$

$$H_2(\mathbf{D}) = -\frac{T}{2}\log(|\mathbf{D}^{\mathrm{T}}\mathbf{D}|).$$

After some calculations, one can show the two terms of (8.38) are

$$\frac{\partial H_1(\widetilde{\boldsymbol{\theta}}, \mathbf{D})}{\partial \mathbf{D}} = \sum_{t-1}^{T} \mathbf{D}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}\{(\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}})(\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}})^{\mathrm{T}}\}(\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1} \quad \text{and} \tag{8.39}$$

$$\frac{\partial H_3(\mathbf{D})}{\partial \mathbf{D}} = -T(\mathbf{D}^{\mathrm{T}})^{+}. \tag{8.40}$$

In order to get the relationship between $\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}}$ and $\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0$, we plug (8.39) and (8.40) back into (8.38), left multiply $T^{-1}\widetilde{\mathbf{D}}^{\mathrm{T}}$ and right multiply $\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}}$ to both sides of the equation. Also we notice $\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}} = \mathcal{X}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}}) + \mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t$ and $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}(\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\boldsymbol{\xi}_t^{\mathrm{T}}\mathbf{D}^0) = \mathbf{D}^{0\mathrm{T}}\boldsymbol{\Sigma}\mathbf{D}^0 = \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0$. Then we can get

$$\begin{aligned}\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0 = \frac{1}{T}\sum_{t=1}^{T}\{&\mathcal{X}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})^{\mathrm{T}}\mathcal{X}_t^{\mathrm{T}} + \mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})^{\mathrm{T}}\mathcal{X}_t^{\mathrm{T}} \\ &+ \mathcal{X}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})\boldsymbol{\xi}_t^{\mathrm{T}}\mathbf{D}^0\}.\end{aligned} \tag{8.41}$$

Using the sub-additivity and sub-multiplicativity of matrix norm, Condition 2(iii), and $\|\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}}\| = O_P(\sqrt{n/T})$ we obtained from Lemma 1, we can show that

$$\begin{aligned}&\|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\| \\ &\quad\le \frac{1}{T}\sum_{t=1}^{T}\|\mathcal{X}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})^{\mathrm{T}}\mathcal{X}_t^{\mathrm{T}}\| + \left\|\frac{2}{T}\sum_{t=1}^{T}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}})^{\mathrm{T}}\mathcal{X}_t^{\mathrm{T}}\right\| \\ &\quad\le \max_t(\|\mathcal{X}_t\mathcal{X}_t^{\mathrm{T}}\|)\|\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}}\|^2 + 2\max_t(\|\mathcal{X}_t\mathcal{X}_t^{\mathrm{T}}\|)\|\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}}\|\left\|\frac{1}{T}\sum_{t=1}^{T}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\right\| \\ &\quad= \max_t(\|\mathcal{X}_t\mathcal{X}_t^{\mathrm{T}}\|)\cdot O_P\left(\frac{n}{T}\right).\end{aligned} \tag{8.42}$$

Now, one can complete the proof of (i) by showing $\max_t\|\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\|$ is upper bounded by some positive constant.

According to the property of matrix norm, we have

$$\max_t \|\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\| \le \Big\{\max_t \sqrt{\|\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\|_1}\Big\}\Big\{\max_s \sqrt{\|\mathcal{X}_s^{\mathrm{T}}\mathcal{X}_s\|_\infty}\Big\}. \tag{8.43}$$

By some calculations, we can show for any $t = 1, \ldots, T$,

$$\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t = \operatorname{diag}(\mathbf{B}_{1t}^{\mathrm{T}}\mathbf{B}_{1t}, \ldots, \mathbf{B}_{nt}^{\mathrm{T}}\mathbf{B}_{nt}),$$

where for $i = 1, \ldots, n$,

$$\mathbf{B}_{it}^{\mathrm{T}}\mathbf{B}_{it} = \begin{pmatrix} 1 & X_{it}^{\mathrm{T}} & \mathbf{0}_p^{\mathrm{T}} \\ X_{it} & X_{it}X_{it}^{\mathrm{T}} & \mathbf{0}_{p\times p} \\ \mathbf{0}_p & \mathbf{0}_{p\times p} & \mathbf{I}_p \end{pmatrix}.$$

From the structure of $\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t$, it is easy to see each row or column of $\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t$ has at most $p+1$ non-zero entries. In addition, according to Condition 2(ii), we have $\max_{i,t}\|X_{it}\| \le C$ for some positive constant C. Hence, consider p is fixed, it is easy to see

$$\begin{aligned} &\max_t \sqrt{\|\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\|_1} \le \sqrt{(p+1)C} \quad \text{and} \\ &\max_t \sqrt{\|\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\|_\infty} \le \sqrt{(p+1)C}. \end{aligned} \tag{8.44}$$

This immediately completes the proof of (i).

The proof of (ii) is readily after the result in (i). Under Condition 2(iii) and let C be a large enough positive constant, we have

$$C^{-1} \le \lambda_{\min}^0 \le \lambda_{\max}^0 \le C.$$

Combining this with (8.42), we have

$$C^{-1} - \|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\| \le \tilde{\lambda}_{\min} \le \tilde{\lambda}_{\max} \le C + \|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\|.$$

As $n^2/T \to 0$, the result in (ii) is shown as one can always find a large enough positive constant C^*, such that

$$C^* > C + \|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\|.$$

Furthermore, it is straightforward to see, $(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}$ is also positive definite and all its eigenvalues are bounded away from 0 and ∞.

To prove (iii), we left multiply $(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}$ to both sides of (8.41) and take the matrix norm of both sides, then we complete the proof of (iii) by showing

$$\|(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{D}^{0\mathrm{T}}\mathbf{D}^0) - \mathbf{I}\| \le \|(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}\|\,\|\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}} - \mathbf{D}^{0\mathrm{T}}\mathbf{D}^0\| = O_p\Big(\frac{n}{T}\Big). \qquad \square$$

LEMMA 3. *Under the conditions of Lemma 2 and $n^2/T \to 0$, we have $\|\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\| = O_P(\frac{1}{\sqrt{T}})$.*

PROOF. Given $\widetilde{\mathbf{D}}$ as the maximum likelihood estimator of $\mathbf{D}$, the first order condition of $\boldsymbol{\theta}$ is

$$
\begin{aligned}
0 &= \frac{\partial L(\boldsymbol{\theta}, \widetilde{\mathbf{D}})}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} = -\frac{1}{2}\sum_{t=1}^{T} \frac{\partial \operatorname{tr}\{(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta})^{\mathrm{T}}(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{Z}_t - \mathcal{X}_t\boldsymbol{\theta})\}}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}} \\
&= \sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}(\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}}).
\end{aligned} \tag{8.45}
$$

By plugging $\mathbf{Z}_t - \mathcal{X}_t\widetilde{\boldsymbol{\theta}} = \mathcal{X}_t(\boldsymbol{\theta}^0 - \widetilde{\boldsymbol{\theta}}) + \mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t$ into (8.45), we get

$$
\begin{aligned}
(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^0) &= \left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}\mathcal{X}_t\right)^{-1}\left[\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\right] \\
&\equiv \mathbf{K}^{-1}\left[\frac{1}{T}\sum_{t=1}^{T} A_t\right],
\end{aligned} \tag{8.46}
$$

where $\mathbf{K} = \frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}\mathcal{X}_t$, and $A_t = \mathcal{X}_t^{\mathrm{T}}(\widetilde{\mathbf{D}}^{\mathrm{T}}\widetilde{\mathbf{D}})^{-1}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t$ for $t = 1, \ldots, T$.

Recall the definition of Ω_i, for $i = 1, \ldots, n$, we can write Ω_i as

$$
\Omega_i = (\mathbf{0}_{p\times\{i(2p+1)-2p\}}\ \mathbf{I}_p\ \mathbf{0}_{p\times\{(n-i)(2p+1)+p\}}). \tag{8.47}
$$

For $i = 1, \ldots, n$, when we left multiply Ω_i to the both sides of (8.46), we have

$$
(\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0) = \Omega_i\mathbf{K}^{-1}\left[\frac{1}{T}\sum_{t=1}^{T} A_t\right]. \tag{8.48}
$$

We first show that all the eigenvalues of $\mathbf{K}^{-1}$ are lower bounded by C^{-1} and upper bounded by C, where C is some large enough positive constant. Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalues of a given square matrix respectively. According to Lemma 2(ii), we can show

$$
\begin{aligned}
C_1^{-1}\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\right) &\le \lambda_{\min}(\mathbf{K}) \le \lambda_{\max}(\mathbf{K}) \\
&\le C_1\lambda_{\max}\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\right),
\end{aligned} \tag{8.49}
$$

where C_1 is a positive constant.

On the one hand, according to Condition 2(ii), we have $\lambda_{\min}(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t) \ge C_2^{-1}$ for some positive constant C_2. So $\lambda_{\min}(\mathbf{K})$ is lower bounded by $C_1^{-1}C_2^{-1}$.

On the other hand, according to (8.42), (8.43) and (8.44), we can upper bounded $\lambda_{\max}(\mathbf{K})$ as follows:

$$
\lambda_{\max}(\mathbf{K}) \le C_1\lambda_{\max}\left(\frac{1}{T}\sum_{t=1}^{T} \mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\right) \le C_1\max_t\|\mathcal{X}_t^{\mathrm{T}}\mathcal{X}_t\| \le C_1C_3(p+1), \tag{8.50}
$$

where C_3 is a positive constant.

Therefore, by choosing a large enough positive constant C, we can show

$$C^{-1} < \lambda_{\min}(\mathbf{K}^{-1}) < \lambda_{\max}(\mathbf{K}^{-1}) < C. \tag{8.51}$$

By taking matrix norm of both sides of (8.48) and using (8.51), we can upper and lower bound $\|\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\|^2$ for $i = 1, \ldots, n$ as

$$\frac{1}{C^2}\left\|\frac{1}{T}\sum_{t=1}^{T}\Omega_i A_t\right\|^2 \le \|\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\|^2 \le C^2\left\|\frac{1}{T}\sum_{t=1}^{T}\Omega_i A_t\right\|^2.$$

Furthermore, according to Lemma 2(ii), we have for $i = 1, \ldots, n$

$$\frac{1}{C^2}\left\|\frac{1}{T}\sum_{t=1}^{T}\Omega_i \mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\right\|^2 \le \left\|\frac{1}{T}\sum_{t=1}^{T}\Omega_i A_t\right\|^2 \le C^2\left\|\frac{1}{T}\sum_{t=1}^{T}\Omega_i \mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\right\|^2.$$

To make the structure of $\mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t$ clear, we denote $\mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t = (\mathbf{d}_{t1}^{\mathrm{T}}, \ldots, \mathbf{d}_{tn}^{\mathrm{T}})^{\mathrm{T}}$, where each $\mathbf{d}_{ti}$ is a $2p+1$ by 1 vector with the following form:

$$\mathbf{d}_{ti} = \begin{pmatrix} \mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it} \\ (\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})X_{it} \\ \Gamma_i \mathbf{f}_t + \boldsymbol{\epsilon}_{it} \end{pmatrix}. \tag{8.52}$$

Recall $\max_{i,t}\|X_{it}\| \le C_3$, then we can show $\max_{i,t,s} X_{it}^{\mathrm{T}}X_{is} \le \max_{i,t}\|X_{it}\|^2 \le C_3^2$. By some calculations, for $i = 1, \ldots, n$, we have

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}\Omega_i \mathcal{X}_t^{\mathrm{T}}\mathbf{D}^{0\mathrm{T}}\boldsymbol{\xi}_t\right\|^2 \\
&\quad= \left\|\frac{1}{T}\sum_{t=1}^{T}(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})X_{it}\right\|^2 \\
&\quad\le \frac{C_3^2}{T}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})^2\right] + \frac{C_3^2}{T^2}\sum_{\substack{t,s=1 \\ t\neq s}}^{T}(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})(\mathbf{f}_s^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{is}) \\
&\quad\equiv I_1 + I_2.
\end{aligned}$$

We first consider I_1. Under Condition 2(i), we have both $E[(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})^2]$ and $\mathrm{Var}((\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})^2)$ are fixed and uniformly upper bounded by a positive constant for all $i = 1, \ldots, n$. Thus, one can easily show I_1 is of order $O_P(1/T)$. We then calculate I_2. Follow the model assumption and Condition 2(i), for $i = 1, \ldots, n$, $t, s = 1, \ldots, T$ and $t > s$, $(\mathbf{f}_t^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{it})(\mathbf{f}_s^{\mathrm{T}}\boldsymbol{\lambda}_i + \varepsilon_{is})$ can be considered as i.i.d. random variables with zero mean and fixed and bounded variance. By the central limit theorem, one can show I_2 is also of order $O_P(1/T)$.

As both I_1 and I_2 are of order $O_P(1/T)$, one can complete the proof by showing

$$C^{*-1}O_P\left(\frac{1}{T}\right) \le \|\widetilde{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_i^0\|^2 \le C^* O_P\left(\frac{1}{T}\right),$$

uniformly for $i = 1, \ldots, n$, where C^* is a large enough positive constant. □

Acknowledgements. The authors thank the Co-Editor, an Associate Editor and two referees for their helpful comments which greatly improved the former version of the paper.

SUPPLEMENTARY MATERIAL

Additional numerical results (DOI: 10.1214/15-AOS1403SUPP; .pdf). We provide additional numerical results for PSID data analysis.

REFERENCES

AHN, S. C. and SCHMIDT, P. (1995). Efficient estimation of models for dynamic panel data. *J. Econometrics* **68** 5–27. MR1345704

ARELLANO, M. (2003). *Panel Data Econometrics*. Oxford Univ. Press, Oxford. MR2060514

BAI, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory* **13** 315–352. MR1455175

BAI, J. and LI, K. (2012). Statistical analysis of factor models of high dimension. *Ann. Statist.* **40** 436–465. MR3014313

BAI, J. and LI, K. (2014). Theory and methods of panel data models with interactive effects. *Ann. Statist.* **42** 142–170. MR3178459

BAI, J. and NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70** 191–221. MR1926259

BALTAGI, B. H. (2005). *Econometric Analysis of Panel Data*. Wiley, Chichester.

BONDELL, H. D. and REICH, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64** 115–123. MR2422825

FRED, A. and JAIN, A. K. (2003). Robust data clustering. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **3** 128–136. IEEE, Madison, WI.

FRIEDMAN, J., HASTIE, T., HÖFLING, H. and TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.* **1** 302–332. MR2415737

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

HSIAO, C. (2003). *Analysis of Panel Data*, 2nd ed. *Econometric Society Monographs* **34**. Cambridge Univ. Press, Cambridge. MR1962511

JIANG, Q., WANG, H., XIA, Y. and JIANG, G. (2013). On a principal varying coefficient model. *J. Amer. Statist. Assoc.* **108** 228–236. MR3174615

KARIYA, T. and KURATA, H. (2004). *Generalized Least Squares*. Wiley, Chichester. MR2120002

KE, Z. T., FAN, J. and WU, Y. (2015). Homogeneity pursuit. *J. Amer. Statist. Assoc.* **110** 175–194. MR3338495

KE, Y., LI, J. and ZHANG, W. (2015). Supplement to "Structure identification in panel data analysis." DOI:10.1214/15-AOS1403SUPP.

LEHMANN, E. L. (1983). *Theory of Point Estimation*. Wiley, New York. MR0702834

PINHEIRO, J. C. and BATES, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

SHEN, X. and HUANG, H.-C. (2010). Grouping pursuit through a regularization solution surface. *J. Amer. Statist. Assoc.* **105** 727–739. MR2724856

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 91–108. MR2136641

YANG, S., YUAN, L., LAI, Y.-C., SHEN, X., WONKA, P. and YE, J. (2012). Feature groupingand selection over an undirected graph. In *Proceedings of the* 18*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 922–930. ACM, New York.

ZHU, Y., SHEN, X. and PAN, W. (2013). Simultaneous grouping pursuit and feature selection over an undirected graph. *J. Amer. Statist. Assoc.* **108** 713–725. MR3174654

Y. KE
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: yuank@princeton.edu

J. LI
DEPARTMENT OF STATISTICS & APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
SINGAPORE 117546
E-MAIL: stalj@nus.edu.sg

W. ZHANG
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF YORK
HESLINGTON
YORK YO10 5DD
UNITED KINGTOM
E-MAIL: wenyang.zhang@york.ac.uk