12-2017

Identifying latent group structures in nonlinear panels

Wuyi WANG
Singapore Management University, wuyi.wang.2013@phdecons.smu.edu.sg

Liangjun SU
Singapore Management University, ljsu@smu.edu.sg

SMU ECONOMICS & STATISTICS

Identifying Latent Group Structures in Nonlinear Panels

Wuyi Wang, Liangjun Su

December 2017

Paper No. 19-2017

Identifying Latent Group Structures in Nonlinear Panels*

Wuyi Wang and Liangjun Su
School of Economics, Singapore Management University

December 16, 2017

Abstract

We propose a procedure to identify latent group structures in nonlinear panel data models where some regression coefficients are heterogeneous across groups but homogeneous within a group and the group number and membership are unknown. To identify the group structures, we consider the order statistics for the preliminary unconstrained consistent estimators of the regression coefficients and translate the problem of classification into the problem of break detection. Then we extend the sequential binary segmentation algorithm of Bai (1997) for break detection from the time series setup to the panel data framework. We demonstrate that our method is able to identify the true latent group structures with probability approaching one and the post-classification estimators are oracle-efficient. The method has the advantage of more convenient implementation compared with some alternative methods, which is a desirable feature in nonlinear panel applications. To improve the finite sample performance, we also consider an alternative version based on the spectral decomposition of certain estimated matrix and link the group identification issue to the community detection problem in the network literature. Simulations show that our method has good finite sample performance. We apply this method to explore how individuals' portfolio choices respond to their financial status and other characteristics using the Netherlands household panel data from year 1993 to 2015, and find three latent groups.

JEL Classification: C33, C38, C51.

Keywords: Binary segmentation algorithm, clustering, community detection, network, oracle estimator, panel structure model, parameter heterogeneity, singular value decomposition.

*The authors thank the conference participants in the 23rd International Panel Data Conference (2017, Thessaloniki), the 13th Symposium on Econometric Theory and Application (SETA 2017, Beijing), and the Advance in Econometrics Conference (2017, Shanghai), and the seminar participants in the Department of Economics at University of York, the Department of Economics at Humboldt-Universitat Zu Berlin, Germany, and the Erasmus School of Economics at Erasmus Universiteit Rotterdam, for their valuable comments. Su acknowledges support from the Singapore Ministry of Education for Academic Research Fund (AcRF) under the Tier-2 grant number MOE2012-T2-2-021 and the funding support provided by the Lee Kong Chian Fund for Excellence. Address correspondence to: Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; Phone: +65 6828 0386; E-mail: ljsu@smu.edu.sg.

1 Introduction

Panel data modeling is one of the most active areas of research in econometrics. By combining individual observations across time, panel data can produce more efficient estimators than pure cross section or time series estimators and allow us to study some problems that are not feasible in the cross section or time series framework. Many advantages of the panel data analysis rest on the parameter homogeneity assumption. Conventional panel data analysis often assumes slope homogeneity to utilize the full power of cross section averaging and make the asymptotic theory easier to derive. Nevertheless, such a homogeneity assumption is frequently called into question and rejected in empirical researches; see Hsiao and Tahmiscioglu (1997), Phillips and Sul (2007), Browning and Carro (2007), Su and Chen (2013), Lu and Su (2017), among others. When the homogeneous slope assumption does not hold, inferences based on it are typically misleading (Hsiao (2014, Chapter 1)). On the other hand, if complete heterogeneity is allowed, the advantages of using panel data can be lost and even the estimation might be impossible. For this reason, more and more researchers consider an intermediate case and study the panel structure model.

In a panel structure model, there exists a subset of parameters that are heterogeneous across groups but homogeneous within a group, and neither the number of groups nor individuals' group membership is known. There are many motivating examples for such a model. In macroeconomics, Phillips and Sul (2007) study the hypothesis of convergence clubs where countries belonging to different groups behave differently; in financial markets, stocks in the same sector share some similar characteristics and behave similarly (Ke, Fan, and Wu (2015)); in labor economics, researchers consider black-white racial differences and classify them into different groups in studying earnings dynamics (Hu (2002)); in economic geography, location is a natural criterion for group classification (Fan, Lv, and Qi (2011); Bester and Hansen (2016)); in international trade, GATT/WTO has uneven impacts on different groups of country-pairs (Subramanian and Wei (2007)). All these examples motivate the use of panel structure models.

To identify the latent group structure is not an easy task. It is computationally infeasible to try all possible combinations of groups, which is a Bell number (Shen and Huang (2010)). Some authors propose to use external variables to determine the group structure; see, e.g., Hu (2002), Subramanian and Wei (2007), and Bester and Hansen (2016). However, this approach may fail for various reasons. For example, it may be impossible to find such an external variable to determine the group structure in empirical studies, and the wrong choice of such a variable can lead to misleading inferences. Several data-driven approaches have been proposed to overcome the shortcomings of reliance on external variables to form groups. One popular approach is based on the K-means algorithm; see Lin and Ng (2012), Sarafidis and Weber (2015), Bonhomme and Manresa (2015), Ando and Bai (2016). The second popular approach is based on the classifier-Lasso (C-Lasso) that has been recently proposed by Su, Shi, and Phillips (2016a, SSP hereafter) and extended in Su and Ju (2017) and Su, Wang, and Jin (2017). In particular, SSP construct a novel C-Lasso procedure where the penalty term is the addition of some multiplicative penalty

terms and show that their method can identify the group structures and estimate the parameters consistently at the same time. In addition, Wang, Phillips, and Su (2017) extend the CARDS algorithm of Ke et al. (2015) to the panel data framework to identify the group structure of slope parameters.

Recently, Ke, Li, and Zhang (2016, KLZ hereafter) borrow the idea of binary segmentation in the structural change literature (e.g., Bai (1997)) and apply it to identify the unobserved group structures in linear panel data models with interactive fixed effects. Let N denote the number of cross sectional units and p the dimension of a parameter vector β_i that is associated with individual i. Let $\boldsymbol{B} = (\beta_1^\top, \ldots, \beta_N^\top)^\top$. KLZ assume that the number of distinct elements in the Np-vector $\boldsymbol{B}$ is given by a finite number, say $\mathcal{N}+1$ in their notation. Based on consistent preliminary estimates $\tilde{\boldsymbol{B}}$ of $\boldsymbol{B}$, they order the elements of $\tilde{\boldsymbol{B}}$ in ascending order and then apply the binary segmentation algorithm sequentially as used in Bai (1997) to identify the group structure and estimate the distinct elements in $\boldsymbol{B}$. Apparently, the setup in KLZ is quite different from the general setup in econometrics where the parameters of interest, β_i as a whole vector, are assumed to be heterogeneous across groups but homogeneous within a group.

Following the lead of Bai (1997) and KLZ, we propose to apply the sequential binary segmentation algorithm (SBSA) to identify the latent group structure on parameter vectors in nonlinear panel data models. In comparison with KLZ, our method is different from theirs in three important ways. First, KLZ consider the classification of scalar coefficients but we consider the classification of parameter vectors. In KLZ's case, there is a natural ordering for their preliminary estimates and they can draw support from the structural change literature where parameters of interest are ordered naturally along the time dimension. In our case, there is no natural order for the estimates of parameter vectors, and fortunately, inspired by the CART-split criterion (Breiman, Friedman, Stone, and Olshen (1984)), we are able to propose a variant of binary segmentation algorithm to classify the vectors. Second, KLZ consider the linear panel data models with interactive fixed effects. They obtain their preliminary estimates by using an EM algorithm and then conduct the binary segmentation based on the ordered preliminary estimates. In contrast, we consider general nonlinear panel data models that contains the linear panel data model as a special case, and apply the modified binary segmentation algorithm on the quasi-maximum likelihood estimates (QMLEs) of the parameter vectors of interest. Third, to determine when the sequential binary segmentation stops, KLZ propose to use the BIC to select a tuning parameter but do not justify the asymptotic validity of information criterion. In contrast, we propose a BIC-type information criterion to determine the number of groups directly and prove that our information criterion can select the number of groups correctly with probability approaching one (w.p.a.1).

In comparison with SSP's C-Lasso method and the K-means algorithm, our method has both pros and cons. First, the K-means algorithm is NP hard and thus computationally demanding. SSP's C-Lasso procedure is not a convex problem but can be transformed into a sequence of convex problems. So the computational burden of SSP's C-Lasso method is not as much as the K-means algorithm but is still quite expensive. In contrast, our SBSA is least computationally demanding

among the three methods. Second, the SSP's C-Lasso need the choice of two tuning parameters, one is used to determine the number of groups, and the other is used for the C-Lasso penalty. Unlike the C-Lasso method but like the K-means algorithm, our binary segmentation algorithm only relies on a single tuning parameter to determine the number of groups via an information criterion. Of course, if the number of groups is known *a priori*, there is no tuning parameter involved in our procedure and the K-means algorithm as well, and one tuning parameter is involved in the C-Lasso procedure. Third, SSP's C-Lasso may leave some individuals unclassified and one has to classify some unclassified individuals after the algorithm based on some distance measure. Like the K-means algorithm, our binary segmentation algorithm forces all individuals to be classified into one of the groups. As SSP argue, leaving some individuals unclassified is not necessarily a bad thing. We also find through our simulations that the preliminary estimates based on some realizations can be rather abnormal when the time dimension T in the panel is not large. In this case, including such abnormal estimates in the algorithm can significantly deteriorate the classification performance. Fourth, in some sense our method can be regarded as a universal method and it works for all panel structure models as long as one can obtain preliminary consistent estimates. The model can be nonstationary panels or panel data models with interactive fixed effects.

In addition, we also allow the presence of common parameters across all individuals. This corresponds to the mixed panel structure model mentioned in SSP (Section 2.7). It is useful when economic theory suggests that some regressors' coefficients are identical across individuals (e.g., Pesaran, Shin, and Smith (1999)). Besides, when a regressor doesn't change over time for many individuals but it is an important factor that must be included in the model, we have no choice but to assume it is homogeneous across individuals. We will illustrate the versatility of the model considered in this paper with examples later.

To enhance the finite sample performance of the SBSA, we also propose an alternative algorithm based on the spectral decomposition of certain symmetric matrix and establish the linkage between the panel structure model and the stochastic block model (SBM) that is widely used for community detection in the network literature (e.g., von Luxburg (2007) and Rohe, Chatterjee, and Yu (2011)). Using a useful variant of the deep Davis-Kahan $\sin\theta$ theorem *a la* Yu, Wang, and Samworth (2015), we are able to show that the individuals' group information is contained in the largest few eigenvectors of such a matrix and it is feasible to conduct SBSA based on such eigenvectors. We also establish the asymptotic distribution theory in this case.

In the application, we study how individuals' portfolio choices are affected by financial assets, non-capital income, retirement status and other factors. Among them, financial assets and non-capital income are modeled to have heterogeneous responses for different individuals. The response variable is the safe asset ratio, which is left censored at 0 and right censored at 1. We use data from the De Nederlandsche Bank (DNB) panel survey. By using the method proposed here, we are able to identify three latent groups. The first group of individuals respond to increasing non-capital income by decreasing the safe assets ratio while the other two groups do the opposite. The increase in financial assets has negative effects for all groups. But the extent is rather different between the

second group and the others. The results are consistent with the general observation that some people tend to invest income on safe assets while others (e.g., risk-loving people) do the contrary.

The rest of the paper is organized as follows. We introduce the latent structure panel data model and the estimation algorithms in Section 2. Asymptotic properties of the algorithm and the final estimators are given in Section 3. In Section 4, we propose an improved algorithm and give its asymptotic properties. In Section 5, we show the finite sample performance of our method by Monte Carlo simulations. In Section 6, we apply our method to study individuals' portfolio choices by using the Netherlands household survey panel data. Section 7 concludes. All proofs are relegated to the appendix.

Notation. For a real matrix (vector) A, we denote its transpose $A^\top$and its Frobenius norm $\|A\|$. When A is symmetric, $\lambda_{\max}(A)$, $\lambda_{\min}(A)$, and $\lambda_j(A)$ denote its largest, smallest, and jth largest eigenvalues, respectively. I_p and $\mathbf{0}_{p\times 1}$ denote the $p \times p$ identity matrix and $p \times 1$ vector of zeros, respectively. $\mathbf{1}\{\cdot\}$ denotes the indicator function. The operators $\xrightarrow{D}$ and $\xrightarrow{P}$ denote convergence in distribution and in probability, respectively.

2 The model and the estimators

In this section we consider the panel structure model and propose a sequential binary segmentation algorithm (SBSA) to estimate the group structures.

2.1 The panel structure model and examples

We consider the general panel data model with latent group structures:

$$y_{it} = g(x_{it}, \varepsilon_{it}; \beta_i, \mu_i, \theta), \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \tag{2.1}$$

where $g(\cdot)$ is a general regression function, x_{it} is a vector of regressors, ε_{it} is the idiosyncratic shock, μ_i is a $r \times 1$ vector of nuisance parameters (e.g., the fixed effects), θ is a $q \times 1$ vector of parameters that is common across individuals, and β_i is a $p \times 1$ vector of parameters whose true values exhibit a group pattern of the general form

$$\beta_i^0 = \sum_{k=1}^{K^0} \alpha_k^0 \cdot \mathbf{1}\left\{i \in G_k^0\right\}.$$

Here $\alpha_k^0 \neq \alpha_l^0$ for any $k \neq l$ and $\mathcal{G}^0 \equiv \{G_1^0, \ldots, G_{K^0}^0\}$ forms a partition of the set $\{1, \ldots, N\}$. We denote the number of individuals in G_k^0 by $N_k \equiv |G_k^0|$, where $|\cdot|$ denotes the cardinality of the set $\cdot$. In this model, the true number of groups K^0 and the group structure $\mathcal{G}^0$ are both unknown.

We denote the minus log-likelihood function of y_{it} conditional on x_{it} and the history of (x_{it}, y_{it}) by $\varphi(w_{it}; \beta_i, \mu_i, \theta)$. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_N)^\top$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{K^0})^\top$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_N)^\top$. The true values of $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\mu}$, and θ are denoted by $\boldsymbol{\beta}^0$, $\boldsymbol{\alpha}^0$, $\boldsymbol{\mu}^0$, and θ^0, respectively. Without any information

about the group structure, we propose to minimize the following objective function

$$L_{NT}(\boldsymbol{\beta}, \boldsymbol{\mu}, \theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi(w_{it}; \beta_i, \mu_i, \theta). \tag{2.2}$$

When the likelihood function is correctly specified, by minimizing the above function we obtain the maximum likelihood estimates (MLEs) $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_N)^\top$, $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_N)^\top$, and $\tilde{\theta}$ of $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, and θ, respectively. Otherwise, they are the quasi-maximum likelihood estimates (QMLEs).

Next, we give some concrete examples for the model in (2.1) and its associated likelihood function in (2.2).

Example 2.1 (Linear panel)**.** We consider two cases.

(i). The standard heterogeneous linear panel data model with individual fixed effects is given by

$$y_{it} = x_{it}^\top \beta_i^0 + \mu_i^0 + \varepsilon_{it}, \tag{2.3}$$

where μ_i is the scalar fixed effect so that $r = 1$, β_i, x_{it}, and ε_{it} are defined as above, and the model does not contain any common parameter of interest so that θ is absent. In this case, we can set $\varphi(w_{it}; \beta_i, \mu_i) = \frac{1}{2}(y_{it} - x_{it}^\top \beta_i - \mu_i)^2$, where $w_{it} = (y_{it}, x_{it}^\top)^\top$.

(ii). Following Pesaran et al. (1999), we can consider a mixed linear panel data model that contains both homogeneous and heterogeneous slope coefficients:

$$y_{it} = x_{1,it}^\top \beta_i^0 + x_{2,it}^\top \theta^0 + \mu_i^0 + \varepsilon_{it},$$

where $x_{it} = (x_{1,it}^\top, x_{2,it}^\top)^\top$ is a $(p+q) \times 1$ vector of regressors, μ_i is the scalar fixed effects, and β_i, θ, and ε_{it} are as defined above. In this case, $\varphi(w_{it}; \beta_i, \mu_i, \theta) = \frac{1}{2}(y_{it} - x_{1,it}^\top \beta_i - x_{2,it}^\top \theta - \mu_i)^2$, where $w_{it} = (y_{it}, x_{1,it}^\top, x_{2,it}^\top)^\top$.

Example 2.2 (Censored panel)**.** The observed response variable y_{it} is subject to two-sided censoring

$$y_{it} = \text{mami}(L, y_{it}^*, R),$$

where the notation $\text{mami}(\cdot)$ is borrowed from Alan et al. (2014) and defined as

$$\text{mami}(L, y, R) = \begin{cases} L & \text{if } y \le L \\ y & \text{if } L < y < R \\ R & \text{if } y \ge R \end{cases}.$$

Clearly, the one-sided censoring is included as a special case by setting $L = -\infty$ or $R = +\infty$ to obtain the right or left censored model. Let $I_{it}^L = \mathbf{1}\{y_{it} = L\}$ and $I_{it}^R = \mathbf{1}\{y_{it} = R\}$. We consider four cases.

(i). The unobserved response variable y_{it}^* is generated as

$$y_{it}^* = x_{it}^\top \beta_i^0 + \mu_i^0 + \varepsilon_{it},$$

and we only observe $\{x_{it}, y_{it}\}$, where $y_{it} = \text{mami}(L, y_{it}^*, R)$, x_{it}, β_i and μ_i are as defined in Example 2.1, ε_{it}'s are independent and identically distributed (i.i.d.) $N\left(0, \sigma^2\right)$. So here the common parameter $\theta = \sigma^2$ and

$$\begin{aligned} -\varphi(w_{it}; \beta_i, \mu_i, \sigma^2) = I_{it}^L \ln \Phi\left((y_{it} - x_{it}^\top \beta_i - \mu_i)/\sigma\right) + I_{it}^R \ln\left(1 - \Phi\left((y_{it} - x_{it}^\top \beta_i - \mu_i)/\sigma\right)\right) \\ + (1 - I_{it}^L - I_{it}^R) \ln\left[\phi\left((y_{it} - x_{it}^\top \beta_i - \mu_i)/\sigma\right)/\sigma\right], \end{aligned} \tag{2.4}$$

where ϕ and Φ denote the probability density function and cumulative distribution function of a standard normal variable, respectively.

(ii). The model in case (i) can be made slightly more general to include a common parameter vector in the regression part:

$$y_{it}^* = x_{1,it}^\top \beta_i^0 + x_{2,it}^\top \theta_2^0 + \mu_i^0 + \varepsilon_{it},$$

where $\theta = (\sigma^2, \theta_2^\top)^\top$ and θ_2 is a $(q-1)$-vector. The QMLE objective function follows directly from (2.4) with $y_{it} - x_{it}^\top \beta_i - \mu_i$ being replaced by $y_{it} - x_{1,it}^\top \beta_i - x_{2,it}^\top \theta_2 - \mu_i$.

(iii). Here the DGP is similar to the first case. The only difference is that ε_{it}'s are i.i.d. $N\left(0, \sigma_i^2\right)$ across t. Then $\mu_i' = (\mu_i, \sigma_i^2)^\top$ plays the role of μ_i in (2.1). The QMLE objective function here is similar to (2.4) but with σ being replaced by σ_i.

(iv). This case is similar to case (ii) except that ε_{it}'s are i.i.d. $N\left(0, \sigma_i^2\right)$ across t. Note that here the individual incidental parameters and common parameters are $(\mu_i, \sigma_i^2)^\top$ and θ, respectively. The QMLE objective function also follows from (2.4) with $y_{it} - x_{it}^\top \beta_i - \mu_i$ and σ being replaced by $y_{it} - x_{1,it}^\top \beta_i - x_{2,it}^\top \theta - \mu_i$ and σ_i, respectively.

Example 2.3 (Binary choice panel)**.** As in Example 2.1, we also consider two cases:

(i). The model is $y_{it} = \mathbf{1}\{x_{it}^\top \beta_i^0 + \mu_i^0 - \varepsilon_{it} \geq 0\}$, where x_{it}, β_i, and μ_i are defined as in Example 2.1 and ε_{it}'s are i.i.d. $N(0,1)$. So in this case, $-\varphi(w_{it}; \beta_i, \mu_i) = y_{it} \ln \Phi(y_{it} - x_{it}^\top \beta_i - \mu_i) + (1 - y_{it}) \ln[1 - \Phi\left(y_{it} - x_{it}^\top \beta_i - \mu_i\right)]$.

(ii). The model is $y_{it} = \mathbf{1}\{x_{1,it}^\top \beta_i^0 + x_{2,it}^\top \theta^0 + \mu_i^0 - \varepsilon_{it} \geq 0\}$. Here, $-\varphi(w_{it}; \beta_i, \mu_i, \theta) = y_{it} \ln \Phi(y_{it} - x_{1,it}^\top \beta_i - x_{2,it}^\top \theta - \mu_i) + (1 - y_{it}) \ln[1 - \Phi(y_{it} - x_{1,it}^\top \beta_i - x_{2,it}^\top \theta - \mu_i)]$.

2.2 Sequential binary segmentation algorithm

The main interest of this paper is to identify the group structure $\mathcal{G}^0$, which contains the information about the number of groups and all individuals' group membership.

To introduce the estimation algorithm, we rewrite the $N \times p$ matrix $\tilde{\boldsymbol{\beta}} \equiv (\tilde{\beta}_1, \ldots, \tilde{\beta}_N)^\top$ as

$$\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_{\cdot 1}, \ldots, \tilde{\boldsymbol{\beta}}_{\cdot p}),$$

where $\tilde{\boldsymbol{\beta}}_{\cdot j}$ denotes the jth column of $\tilde{\boldsymbol{\beta}}$ for $j = 1, \ldots, p$. Let $\beta^0_{i,j}$, $\alpha^0_{k,j}$ and $\tilde{\beta}_{i,j}$ denote the jth element of β^0_i, α^0_k and $\tilde{\beta}_i$, respectively, for $j = 1, \ldots, p$. We sort the N elements of $\tilde{\boldsymbol{\beta}}_{\cdot j}$ in ascending order and denote the order statistics by

$$\tilde{\beta}_{\pi_j(1),j} \leq \tilde{\beta}_{\pi_j(2),j} \leq \cdots \leq \tilde{\beta}_{\pi_j(N),j}, \tag{2.5}$$

where $\{\pi_j(1), \ldots, \pi_j(N)\}$ is a permutation of $\{1, \ldots, N\}$ that is implicitly determined by the order relation in (2.5). Let

$$\mathcal{S}_{i,l}(j) \equiv \{\tilde{\beta}_{\pi_j(i),j}, \tilde{\beta}_{\pi_j(i+1),j}, \ldots, \tilde{\beta}_{\pi_j(l),j}\}$$

for $1 \leq i < l \leq N$.

Fix $j \in \{1, \ldots, p\}$. Intuitively speaking, if the $\beta^0_{i,j}$'s are not identical across i for some j, then finding the homogeneity among $\beta^0_{i,j}$'s is equivalent to finding the "break points" among the ordered version of $\beta^0_{i,j}$'s. When $\tilde{\beta}_{i,j}$'s are consistent estimates of $\beta^0_{i,j}$'s, we expect the "break points" in the ordered $\beta^0_{i,j}$'s will be carried upon to the ordered $\tilde{\beta}_{i,j}$'s. Consequently, we can apply the binary segmentation algorithm sequentially to detect all breaks among the ordered $\beta^0_{i,j}$'s. For example, suppose $K^0 = 3$, $\alpha^0_{1,j} < \alpha^0_{2,j} < \alpha^0_{3,j}$, and N_1 (resp. N_2 and $N - N_1 - N_2$) $\beta^0_{i,j}$'s take value $\alpha^0_{1,j}$ (resp. $\alpha^0_{2,j}$ and $\alpha^0_{3,j}$). Then we expect to see two break points in the sequence $\mathcal{S}_{1,N}(j) = \{\tilde{\beta}_{\pi_j(1),j}, \tilde{\beta}_{\pi_j(2),j}, \ldots, \tilde{\beta}_{\pi_j(N),j}\}$ in large samples that are given by N_1 and $N_1 + N_2$. This is simply because when the sample size is sufficiently large, all elements in the subsamples $\mathcal{S}_{1,N_1}(j)$, $\mathcal{S}_{N_1+1,N_1+N_2}(j)$, and $\mathcal{S}_{N_1+N_2+1,N}(j)$ have the probability limits $\alpha^0_{1,j}$, $\alpha^0_{2,j}$, and $\alpha^0_{3,j}$, respectively. We will show that w.p.a.1, we can identify the two break points N_1 and $N_1 + N_2$ based on the ranking relationship in (2.5) provided that $\alpha^0_{1,j}$, $\alpha^0_{2,j}$, and $\alpha^0_{3,j}$ are distinct from each other.

Complications arise here because it is possible for all $j \in \{1, \ldots, p\}$, $\alpha^0_{1,j}, \ldots,$ and $\alpha^0_{K^0,j}$ are not all distinct from each other and K^0 is typically unknown. For this reason, we have to allow the possibility that $\{\alpha^0_{k,j}, k = 1, \ldots, K^0\}$ are not all distinct from each other for all j and the possibility that $\alpha^0_{1,j} = \cdots = \alpha^0_{K^0,j}$ for some j. We achieve the identification of all K^0 groups based on the key observation that the sample variance of the subsample $\mathcal{S}_{i,l}(j)$ behaves quite differently depending on whether $\beta^0_{\pi_j(i),j}$ is the same as $\beta^0_{\pi_j(l),j}$. If $\beta^0_{\pi_j(i),j} = \beta^0_{\pi_j(i+1),j} = \cdots = \beta^0_{\pi_j(l),j}$, then the sample variance of $\mathcal{S}_{i,l}(j)$ is proportional to T^{-1} when the preliminary estimates $\tilde{\beta}_i$ are all $\sqrt{T}$-consistent; on the other hand, if there is a break between i and l such that $\beta^0_{\pi_j(i),j} < \beta^0_{\pi_j(l),j}$, then the sample variance of $\mathcal{S}_{i,l}(j)$ will be bounded away from zero. This motivates us to choose regressor index j such that $\tilde{\beta}_{i,j}$'s has the largest variance in the investigated segment (i, l) to detect a possible break point.

Let

$$\bar{\beta}_{i,l}(j) = \frac{1}{l-i+1} \sum_{i'=i}^{l} \tilde{\beta}_{\pi_j(i'),j} \quad \text{and } \hat{V}^0_{i,l}(j) \equiv \frac{1}{l-i} \sum_{i'=i}^{l} [\tilde{\beta}_{\pi_j(i'),j} - \bar{\beta}_{i,l}(j)]^2$$

denote the sample mean and variance of the subsample $\mathcal{S}_{i,l}(j)$, respectively. Let $\hat{\sigma}_i^2(j)$ denote a consistent estimator of the asymptotic variance $\mathrm{Var}(\sqrt{T}\tilde{\beta}_{\pi_j(i),j})$. Let $\hat{V}_{i,l}(j) \equiv \hat{V}_{i,l}^0(j)/\bar{\sigma}_{i,l}^2(j)$ where $\bar{\sigma}_{i,l}^2(j) = \frac{1}{l-i+1}\sum_{i'=i}^{l}\hat{\sigma}_{i'}^2(j)$. Define

$$\hat{S}_{i,l}(j,m) = \frac{1}{l-i+1}\left\{\sum_{i'=i}^{m}\left[\tilde{\beta}_{\pi_j(i'),j} - \bar{\beta}_{i,m}(j)\right]^2 + \sum_{i'=m+1}^{l}\left[\tilde{\beta}_{\pi_j(i'),j} - \bar{\beta}_{m+1,l}(j)\right]^2\right\}. \tag{2.6}$$

Since K^0 is typically unknown, we have to pick up a large enough number $K^{\max}$ such that $1 \leq K^0 \leq K^{\max}$. Let K denote a generic number of groups. We propose to adopt the following SBSA to estimate $\mathcal{G}^0$.

Sequential Binary Segmentation Algorithm 1 (SBSA 1)[1]

1. Let $K \in [1, K^{\max}]$. When $K = 1$, there is only one group, i.e., slope coefficients β_i's are actually homogeneous. In this case, the estimated group $\hat{G}_1(1) = \{1, \ldots, N\}$.

2. When $K = 2$, let $\hat{\jmath}_1 = \arg\max_{1\leq j\leq p}\hat{V}_{1,N}(j)$. Given $\hat{\jmath}_1$, we solve the following minimization problem
$$\hat{m}_1 \equiv \underset{1\leq m<N}{\arg\min}\,\hat{S}_{1,N}(\hat{\jmath}_1, m).$$
Now we have two segments – $\hat{G}_1(2) = \mathcal{S}_{1,\hat{m}_1}(\hat{\jmath}_1)$ and $\hat{G}_2(2) = \mathcal{S}_{\hat{m}_1+1,N}(\hat{\jmath}_1)$.

3. When $K \geq 3$, we use $\hat{m}_1, \ldots, \hat{m}_{K-2}$ denote the break points detected in the previous steps such that $\hat{m}_1 < \cdots < \hat{m}_{K-2}$ perhaps after relabeling the $K-2$ break points that have been detected so far. Define
$$\begin{aligned}\hat{\jmath}_{K-1} &\equiv \underset{1\leq j\leq p}{\arg\max}\sum_{k=1}^{K-1}\hat{V}_{\hat{m}_{k-1}+1,\hat{m}_k}(j),\\ \hat{m}_{K-1}(k) &\equiv \underset{\hat{m}_{k-1}+1\leq m<\hat{m}_k}{\arg\min}\,\hat{S}_{\hat{m}_{k-1}+1,\hat{m}_k}(\hat{\jmath}_{K-1}, m) \text{ for } k = 1, \ldots, K-1,\end{aligned}$$
where $\hat{m}_0 = 0$, $\hat{m}_{K-1} = N$, and we suppress the dependence of $\hat{m}_{K-1}(k)$ on $\hat{\jmath}_{K-1}$. Then $\hat{m}_{K-1}(k)$ divides $\hat{G}_k(K-1)$ into two subsegments, which are labeled as $\hat{G}_{k1}(K-1)$ and $\hat{G}_{k2}(K-1)$ respectively. Calculate for $k = 1, \ldots, K-1$,
$$\begin{aligned}\hat{S}_{K-1}(k) \equiv &\sum_{i\in\hat{G}_{k1}(K-1)}\left[\tilde{\beta}_{i,\hat{\jmath}_{K-1}} - \bar{\beta}_{\hat{G}_{k1}(K-1)}(\hat{\jmath}_{K-1})\right]^2 + \sum_{i\in\hat{G}_{k2}(K-1)}\left[\tilde{\beta}_{i,\hat{\jmath}_{K-1}} - \bar{\beta}_{\hat{G}_{k2}(K-1)}(\hat{\jmath}_{K-1})\right]^2\\ &+ \sum_{1\leq l\leq K-1, l\neq k}\sum_{i\in\hat{G}_l(K-1)}\left[\tilde{\beta}_{i,\hat{\jmath}_{K-1}} - \bar{\beta}_{\hat{G}_l(K-1)}(\hat{\jmath}_{K-1})\right]^2,\end{aligned}$$

[1] A major difference between our algorithm and that of KLZ is that KLZ specify a tuning parameter δ that is compared with something similar to our $S_{1,N}(j,m)$ to determine when one should stop the algorithm. Even though they propose to use the BIC to choose δ, there is no asymptotic justification for this. In contrast, we propose to use an information criterion to determine the number of groups directly and justify its asymptotic validity. Admittedly, $K^{\max}$ plays the role of δ in KLZ but our result is insensitive to its choice.

where, e.g., $\bar{\beta}_{\hat{G}_{k1}(K-1)}(\hat{\jmath}_{K-1}) = |\hat{G}_{k1}(K-1)|^{-1} \sum_{i \in \hat{G}_{k1}(K-1)} \tilde{\beta}_{i,\hat{\jmath}_{K-1}}$. Let

$$\hat{k} = \underset{1 \le k \le K-1}{\arg\min}\ \hat{S}_{K-1}(k).$$

We now obtain the $K-1$ break points and the K segments given by $\{\hat{m}_1, \ldots, \hat{m}_{K-2}, \hat{m}_{K-1}(\hat{k})\}$ and $\{\hat{G}_1(K-1), \ldots, \hat{G}_{\hat{k}-1}(K-1), \hat{G}_{\hat{k}1}(K-1), \hat{G}_{\hat{k}2}(K-1), \hat{G}_{\hat{k}+1}(K-1), \ldots, \hat{G}_{K-1}(K)\}$, respectively. Relabel these $K-1$ break points as $\{\hat{m}_1, \ldots, \hat{m}_{K-1}\}$ such that $\hat{m}_1 < \hat{m}_2 < \cdots < \hat{m}_{K-1}$, and the corresponding K groups as $\{\hat{G}_1(K), \hat{G}_2(K), \ldots, \hat{G}_K(K)\}$.

4. Repeat the last step until $K = K^{\max}$.

Of course if K^0 is known *a priori*, we can set $K^{\max} = K^0$. At the end of the SBSA 1, we obtain the $\hat{\mathcal{G}}(K^0) \equiv \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{K^0}\}$ as the estimates of the true group structure $\mathcal{G}^0$. Otherwise, we need first to estimate K^0 before we obtain the final estimate of $\mathcal{G}^0$. See the next subsection.

2.3 The estimation of the model parameters

Let $\hat{\mathcal{G}}(K) \equiv \{\hat{G}_1(K), \hat{G}_2(K), \ldots, \hat{G}_K(K)\}$. Given the estimated group structure $\hat{\mathcal{G}}(K)$ for $K \in [1, K^{\max}]$, we propose to estimate the model parameters by minimizing

$$L_{NT}(\boldsymbol{\beta}, \boldsymbol{\mu}, \theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi(w_{it}; \beta_i, \mu_i, \theta)$$
$$\text{s.t. } \beta_i = \alpha_k \text{ for } i \in \hat{G}_k(K) \text{ and } k = 1, \ldots, K. \tag{2.7}$$

Let $\hat{\boldsymbol{\beta}}(K), \hat{\boldsymbol{\mu}}(K), \hat{\theta}(K)$, and $\hat{\alpha}(K)$ denote the solution to the above minimization problem, where $\hat{\boldsymbol{\beta}}(K) = (\hat{\beta}_1(K), \ldots, \hat{\beta}_N(K))^\top$, $\hat{\boldsymbol{\mu}}(K) = (\hat{\mu}_1(K), \ldots, \hat{\mu}_N(K))^\top$, $\hat{\alpha}(K) = (\hat{\alpha}_1(K), \ldots, \hat{\alpha}_K(K))^\top$, and $\hat{\alpha}_k(K)$ is the estimate of the group-specific parameter vector α_k. We propose to select K to minimize the following BIC-type information criterion

$$\mathrm{IC}_1(K) = 2L_{NT}(\hat{\boldsymbol{\beta}}(K), \hat{\boldsymbol{\mu}}(K), \hat{\theta}(K)) + pK \cdot \rho_{NT}, \tag{2.8}$$

where ρ_{NT} is a tuning parameter that plays the role of $\ln(NT)/(NT)$ in the use of BIC in the panel setup. Let

$$\hat{K} \equiv \underset{1 \le K \le K^{\max}}{\arg\min}\ \mathrm{IC}_1(K) \text{ and } \hat{\mathcal{G}} \equiv \hat{\mathcal{G}}(\hat{K}) \equiv \{\hat{G}_1(\hat{K}), \hat{G}_2(\hat{K}), \ldots, \hat{G}_{\hat{K}}(\hat{K})\}. \tag{2.9}$$

We will show that

$$P(\hat{K} = K^0) \to 1 \text{ and } P(\hat{\mathcal{G}} = \mathcal{G}^0) \text{ as } (N, T) \to \infty.$$

Given $\hat{K}$ and $\hat{\mathcal{G}}$, we consider the constrained minimization problem in (2.7) with K being replaced by $\hat{K}$ and obtain the final estimate of $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$, and θ as

$$\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\hat{K}) = (\hat{\beta}_1(\hat{K}), \ldots, \hat{\beta}_N(\hat{K}))^\top, \qquad \hat{\boldsymbol{\mu}} \equiv \hat{\boldsymbol{\mu}}(\hat{K}) = (\hat{\mu}_1(\hat{K}), \ldots, \hat{\mu}_N(\hat{K}))^\top,$$
$$\hat{\boldsymbol{\alpha}} \equiv \hat{\boldsymbol{\alpha}}(\hat{K}) = (\hat{\alpha}_1(\hat{K}), \ldots, \hat{\alpha}_{\hat{K}}(\hat{K}))^\top, \qquad \hat{\theta} \equiv \hat{\theta}(\hat{K}).$$

Note that these estimates can be obtained via the standard profile maximum likelihood method once we have the estimated group structure $\hat{\mathcal{G}}$. That is, $\hat{\boldsymbol{\alpha}}$ and $\hat{\theta}$ can be obtained as the minimizer of the following objective function

$$\hat{Q}_{NT}(\boldsymbol{\alpha}, \theta) = \frac{1}{NT} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_k(\hat{K})} \sum_{t=1}^{T} \varphi(w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k, \theta), \theta), \tag{2.10}$$

where $\hat{\mu}_i(\alpha_k, \theta) = \arg\min_{\mu_i} \frac{1}{T} \sum_{t=1}^{T} \varphi(w_{it}; \alpha_k, \mu_i, \theta)$ for $i \in \hat{G}_k(\hat{K})$ and $k = 1, \ldots, \hat{K}$. We will study the asymptotic properties of $\hat{\boldsymbol{\alpha}}$ and $\hat{\theta}$ in the next section.

3 Asymptotic properties

In this section, we first study the consistency of the preliminary estimates and then study the asymptotic properties of our estimates of the group structure and other model parameters.

3.1 Consistency of the preliminary estimates

Let $\gamma_i = (\beta_i^\top, \mu_i^\top)^\top$, $\varsigma_i = (\gamma_i^\top, \theta^\top)^\top$, $\gamma_i^0 = (\beta_i^{0\top}, \mu_i^{0\top})^\top$, and $\varsigma_i^0 = (\gamma_i^{0\top}, \theta^{0\top})^\top$. Following the literature on nonlinear panels (e.g., Hahn and Newey (2004), Hahn and Kuersteiner (2011), and SSP), we consider the profile log-likelihood function

$$Q_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi\left(w_{it}; \tilde{\gamma}_i(\theta), \theta\right), \tag{3.1}$$

where $\tilde{\gamma}_i(\theta) = \arg\min_{\gamma_i} \frac{1}{T} \sum_{t=1}^{T} \varphi(w_{it}; \gamma_i, \theta)$. Let $\tilde{\theta} = \arg\min_\theta Q_{NT}(\theta)$ and $\tilde{\gamma}_i = \tilde{\gamma}_i(\tilde{\theta}) = (\tilde{\beta}_i^\top, \tilde{\mu}^\top)^\top$. Let

$$\gamma_i(\theta) \equiv \arg\min_{\gamma_i} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\varphi(w_{it}; \gamma_i, \theta)].$$

Note that $\gamma_i^0 = \gamma_i(\theta^0)$ for $i = 1, \ldots, N$.

Let $Z(w_{it}; \gamma_i, \theta) \equiv \partial \varphi(w_{it}; \gamma_i, \theta)/\partial \gamma_i$ and $W(w_{it}; \gamma_i, \theta) \equiv \partial \varphi(w_{it}; \gamma_i, \theta)/\partial \theta$. Let Z^{γ_i} denote the first derivative of Z with respect to $\gamma_i^\top$. Define W^{γ_i} and W^θ similarly. Define

$$H_{i,\gamma\gamma}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[Z^{\gamma_i}(w_{it}; \gamma_i(\theta), \theta)\right] \text{ and } H_{i,\theta\theta}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[W_{it}^{\theta}(\theta) + W_{it}^{\mu_i}(\theta) \frac{\partial \gamma_i(\theta)}{\partial \theta^\top}\right],$$

where $W_{it}^{\theta}(\theta) = W_i^{\theta}(w_{it}; \gamma_i(\theta), \theta)$ and $W_{it}^{\mu_i}(\theta) = W_i^{\mu_i}(w_{it}; \gamma_i(\theta), \theta)$. For notational simplicity, let $\max_i$ and $\max_{i,t}$ abbreviate $\max_{1 \le i \le N}$ and $\max_{1 \le i \le N, 1 \le t \le T}$, respectively, and similarly for $\min_i$ and $\min_{i,t}$.

To state the first main result, we make the following assumptions.

Assumption A1 (i) For each i, $\{w_{it}, t \ge 1\}$ is stationary strong mixing with mixing coefficient $\alpha_i(\cdot)$. Let $\alpha(\cdot) \equiv \max_i \alpha_i(\cdot)$ satisfies $\alpha(s) \le c_\alpha \rho^s$ for some $c_\alpha > 0$ and $\rho \in (0, 1)$. $\{w_{it}\}$ are independent across i.

(ii) For any $\eta > 0$, there exists a constant $\epsilon > 0$ such that $\min_i\{\min_{\varsigma_i:\|\varsigma_i-\varsigma_i^0\|>\eta} \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\varphi(w_{it};\varsigma_i) - \varphi(w_{it};\varsigma_i^0)]\} > \epsilon$ and $\inf_{\theta:\|\theta-\theta^0\|>\eta} \frac{1}{N}\sum_{i=1}^N \left[\Psi_i\left(\gamma_i\left(\theta\right),\theta\right) - \Psi_i(\gamma_i(\theta^0),\theta^0)\right] > \epsilon$, where $\Psi_i\left(\gamma_i,\theta\right) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\varphi(w_{it};\gamma_i,\theta)]$.

(iii) Let Υ and Θ denote the parameter space for ς_i and θ, respectively. Υ is compact and convex and the true value ς_i^0 lies in the interior of Υ for all $i = 1,\ldots,N$.

(iv) For a $(p+r+q)\times 1$ vector $d = (d_1,\ldots,d_{p+r+q})^\top \in \mathbb{N}^{p+r+q}$, we let $|d|$ denote $\sum_{j=1}^{p+r+q} d_j$. Let $D^d\varphi_{it}(w_{it};\varsigma_i) \equiv \partial^{|d|}\varphi_{it}(w_{it};\varsigma_i)/\partial^{d_1}\varsigma_{i,1}\cdots\partial^{d_{p+r+q}}\varsigma_{i,p+r+q}$, where $\varsigma_{i,j}$ denotes the jth element of ς_i. There is a non-negative real function $M(\cdot)$ such that $\sup_{\varsigma_i\in\Upsilon}\|D^d\varphi_{it}(w_{it};\varsigma_i)\| \le M(w_{it})$ and $\|D^d\varphi_{it}(w_{it};\varsigma_i) - D^d\varphi_{it}(w_{it};\varsigma_i')\| \le M(w_{it})\|\varsigma_i - \varsigma_i'\|$ for all ς_i, $\varsigma_i' \in \Upsilon$ and $|d| \le 3$, and $\max_i \mathbb{E}|M(w_{it})|^\kappa \le c_M$ for some $c_M < \infty$ and $\kappa \ge 6$.

(v) There exists a finite constant $c_H > 0$ such that $\min_i \inf_{\theta\in\Theta} \lambda_{\min}\left(H_{i,\gamma\gamma}(\theta)\right) \ge c_H$ and $\min_i \lambda_{\min}\left(H_{i,\theta\theta}(\theta^0)\right) \ge c_H$.

(vi) $NT^{1-\kappa/2} \to c \in [0,\infty)$ as $(N,T)\to\infty$.

Assumptions A1(i)–(v) parallel Assumptions A1(i)–(v) in SSP. Assumption A1(i) imposes that w_{it}'s are independent across individuals and strong mixing over time. This condition is commonly assumed in the nonlinear panel literature; see, e.g., Hahn and Kuersteiner (2011) and SSP. The stationarity condition is not necessary; it is assumed only for the purpose of simplifying the notation in the proofs of some asymptotic results in the appendix. Assumption A1(ii) imposes the identification condition for the common parameter θ. Assumption A1(iii) requires $\{\varsigma_i\}$ take values in the same bounded and closed subset of $\mathbb{R}^{p+r+q}$. Assumption A1(iv) requires $\varphi(\cdot)$ and its partial derivatives up to the third order are sufficiently smooth and satisfying some moment conditions. Assumption A1(v) assumes that the Hessian matrices $H_{i,\gamma\gamma}(\theta)$ and $H_{i,\theta\theta}(\theta^0)$ have eigenvalues that are bounded away from zero. Assumption A1(vi) restricts that N can not diverge to infinity too fast relative to T. In particular, we allow $N/T^2 \to c \in [0,\infty)$ if $\kappa = 6$.

The following theorem establishes the consistency of the preliminary estimates $\tilde\theta$ and $\tilde\gamma_i$.

Theorem 3.1 (Consistency of preliminary estimators)**.** *Suppose Assumption A1 holds. Then (i)* $\tilde\theta - \theta^0 = O_P\left(T^{-1/2}\right)$, *(ii)* $\tilde\gamma_i - \gamma_i^0 = O_P\left(T^{-1/2}\right)$, *(iii)* $\max_{1\le i\le N}\left\|\tilde\gamma_i - \gamma_i^0\right\| = O_P(T^{-1/2}\left(\ln T\right)^3)$, *and (iv)* $\frac{1}{N}\sum_{i=1}^N\left\|\tilde\gamma_i - \gamma_i^0\right\|^2 = O_P(T^{-1})$.

The proof of the above theorem is rather complicated and relegated to the appendix. The rate in Theorem 3.1(iii) is not optimal. In fact, following Su, Shi, and Phillips (2016b, SSPb hereafter) we can establish that $P(\max_{1\le i\le N}\left\|\tilde\gamma_i - \gamma_i^0\right\| \ge CT^{-1/2}\left(\ln T\right)^3) = o(N^{-1})$ for some large positive constant C. We can obtain a slightly tighter probability order for $\max_{1\le i\le N}\left\|\tilde\gamma_i - \gamma_i^0\right\|$ when we do not restrict the above tail probability to be $o(N^{-1})$.

3.2 Consistency of classification

To study the classification consistency, we introduce some additional notation. Let $\mathcal{G}(K) = \{G_1\left(K\right), G_2\left(K\right),\ldots,G_K\left(K\right)\}$ be an arbitrary partition of $\{1,\ldots,N\}$ where $|G_k\left(K\right)| \ge 1$ for

$k = 1, \ldots, K$. Define $\hat{\sigma}^2_{\mathcal{G}(K)} = 2(NT)^{-1} \sum_{k=1}^{K} \sum_{i \in G_k} \sum_{t=1}^{T} \varphi(w_{it}; \check{\beta}_i(K), \check{\mu}_i(K), \check{\theta}(K))$, where $\check{\beta}_i(K), \check{\mu}_i(K)$, and $\check{\theta}(K)$ solve the constrained problem in (2.7) with $\{\hat{G}_k\}$ being replaced by $\{G_k(K)\}$.

We add two assumptions.

Assumption A2 (i) There exists a constant $c_L > 0$ such that slopes $\min_{1 \le k < k' \le K^0} \|\alpha^0_k - \alpha^0_{k'}\| > c_L$.

(ii) The number of groups K^0 is fixed. $N_k/N \to \tau_k \in (0,1)$ as $N \to \infty$ for $k = 1, \ldots, K^0$.

Assumption A3 (i) $N^{1/2}(\ln N)^9/T \to 0$ as $(N,T) \to \infty$.

(ii) As $(N,T) \to \infty$, $\min_{1 \le K < K^0} \min_{\mathcal{G}(K)} \hat{\sigma}^2_{\mathcal{G}(K)} \xrightarrow{P} \bar{\sigma}^2 > \sigma^2_0$, where $\sigma^2_0 \equiv \lim_{(N,T)\to\infty} 2(NT)^{-1} \sum_{k=1}^{K^0} \sum_{i \in G^0_k} \sum_{t=1}^{T} \mathbb{E}\varphi(w_{it}; \alpha^0_k, \mu^0_i, \theta^0)$.

(iii) $\rho_{NT} \to 0$ as $(N,T) \to \infty$ and $T\rho_{NT} \to \infty$ as $(N,T) \to \infty$.

Assumption A2(i)–(ii) is commonly assumed in the literature on panel structure models; see, e.g., Bonhomme and Manresa (2015) and SSP. Assumption A2(i) requires the minimum distance between the group-specific parameters are bounded away from zero. At the cost of more complicated arguments, we can allow $\min_{1 \le k < k' \le K^0} \|\alpha^0_k - \alpha^0_{k'}\|$ to shrink to zero at a rate slower than $T^{-1/2}(\ln T)^3$. But in practice, when the group-specific parameters are not sufficiently separated from each other, it is hard to estimate the group structure accurately with any finite period of time series observations. Assumption A2(ii) requires each group has an nonnegligible ratio of members asymptotically. Assumption A3(i) strengthens the condition in Assumption A1(vi) to ensure that the estimation error from the preliminary estimates does not play a role in the determination of the number of groups and the asymptotic distribution of our final estimators. Note that unlike KLZ who require $(N \ln N)^2/T \to 0$, we allow N to diverge to infinity at a faster rate than T. A reason for such a big distinction is that we explicitly evaluate the smaller order terms in the differences of the objective functions in the proof of Theorem 3.2 below while KLZ only apply a rough probability bound to control them. Assumption A3(ii)–(iii) imposes some typical conditions to ensure both over-grouped and under-grouped panel structure models are ruled out. In particular, Assumption A3(ii) ensures that for all under-fitted models, the mean square errors would be asymptotically greater than σ^2_0.

The following theorem indicates that we can estimate the true group structure $\mathcal{G}^0$ in the case of known number of groups.

Theorem 3.2 (Classification consistency)**.** *Suppose Assumptions A1–A2 hold. Suppose the true number of groups is known to be K^0. Let $\hat{\mathcal{G}}(K^0) = \{\hat{G}_1(K^0), \ldots, \hat{G}_{K^0}(K^0)\}$ be the estimated group structure based on the SBSA 1. Then $P(\hat{\mathcal{G}}(K^0) = \mathcal{G}^0) \to 1$ as $(N,T) \to \infty$.*

Theorem 3.2 shows that when the true number of groups (K^0) is known, we can estimate the true group structure $\mathcal{G}^0$ correctly w.p.a.1. The proof of Theorem 3.2 relies on the result in Theorem 3.1 but is quite involved.

Nevertheless, K^0 is typically unknown in practice. In this case we need to rely on the information criterion in (2.8) to determine the number of groups. The following theorem establishes the consistency of the information criterion.

Theorem 3.3 (Consistency of the information criterion)**.** *Suppose Assumptions A1–A3 hold. Let $\hat{K}$ be as defined in (2.9). Then $P(\hat{K} = K^0) \to 1$ as $(N,T) \to \infty$.*

That is, we can consistently estimate the number of groups in practice. By using $\hat{K}$ in place of K^0, we can estimate the true group structure $\mathcal{G}^0$ w.p.a.1 by Theorems 3.2 and 3.3.

Note that the last condition in Assumption A3(iii) imposes that $T\rho_{NT} \to \infty$ as $(N,T) \to \infty$ so that ρ_{NT} can only converge to zero at a speed slower than T^{-1}. This is simply due to the fact that the heterogeneous incidental parameters μ_i's in the model can only be estimated at the slow $T^{-1/2}$ convergence rate. For linear panel data models where μ_i is an additive fixed effect, it can be eliminated through the within-group transformation and does not affect the convergence rate of the estimator of the error variance in the model. In this case, we can easily relax Assumption A3(iii) to

Assumption A3 (iii*) $\rho_{NT} \to 0$ as $(N,T) \to \infty$ and $(NT + T^2)\rho_{NT} \to \infty$ as $(N,T) \to \infty$.

If the constrained estimates of β_i's in (2.7) for the linear model are bias corrected. The above condition can be further relaxed to

Assumption A3 (iii**) $\rho_{NT} \to 0$ as $(N,T) \to \infty$ and $NT\rho_{NT} \to \infty$ as $(N,T) \to \infty$.

An implication for this is that the usual BIC information criterion ($\rho_{NT} = \ln(NT)/(NT)$) is also working in our framework when the model is linear and the estimators are bias-corrected.

3.3 Asymptotic distribution

In this section, we study the asymptotic distributions of $\hat{\alpha}_k$'s and $\hat{\theta}$. Recall that $W(w_{it};\beta_i,\mu_i,,\theta) \equiv \partial\varphi(w_{it};\beta_i,\mu_i,\theta)/\partial\theta$. Let $U(w_{it};\beta_i,\mu_i,\theta) = \partial\varphi(w_{it};\beta_i,\mu_i,\theta)/\partial\beta_i$ and $V(w_{it};\beta_i,\mu_i,\theta) \equiv \partial\varphi(w_{it};\beta_i,\mu_i,\theta)/\partial\mu_i$. Let U_j denotes the jth element in U, and similarly for V_j and W_j. Let U^β denote the derivative of U with respect to $\beta^\top$. Define U^μ, V^β, V^μ, V^θ, W^μ and W^θ analogously. For notational simplicity, let $U_{it} \equiv U(w_{it};\beta_i^0,\mu_i^0,\theta^0)$, and similarly for V_{it}, W_{it}, U_{it}^μ, V_{it}^β, V_{it}^μ, V_{it}^θ, W_{it}^μ and W_{it}^θ. Let $U_{it,j}^\mu \equiv \partial U_j(w_{it};\beta_i^0,\mu_i^0,\theta^0)/\partial\mu_i^\top$, $U_{it,j}^{\mu\mu} \equiv \partial^2 U_j(w_{it};\beta_i^0,\mu_i^0,\theta^0)/\partial\mu_i\partial\mu_i^\top$, and similarly for $W_{it,j}^\mu$, $V_{it,j}^{\mu\mu}$ and $W_{it,j}^{\mu\mu}$. Define

$$S_{iU} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}(U_{it}^\mu),\ S_{iV} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}(V_{it}^\mu),\ S_{iW} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}(W_{it}^\mu),$$

$$S_{iU2,j} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}(U_{it,j}^{\mu\mu}),\ S_{iV2,j} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}(V_{it,j}^{\mu\mu}),\ S_{iW2,j} \equiv \frac{1}{T}\sum_{t=1}^T \mathbb{E}(W_{it,j}^{\mu\mu}),$$

$$\mathbb{U}_{it} \equiv U_{it} - S_{iU}S_{iV}^{-1}V_{it},\ \mathbb{U}_{it}^\mu \equiv U_{it}^\mu - S_{iU}S_{iV}^{-1}V_{it}^\mu,\ \mathbb{W}_{it} \equiv W_{it} - S_{iW}S_{iV}^{-1}V_{it},\ \mathbb{W}_{it}^\mu \equiv W_{it}^\mu - S_{iW}S_{iV}^{-1}V_{it}^\mu,$$

$$\Omega_{iT,\beta\beta} \equiv \frac{1}{T}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}(\mathbb{U}_{is}\mathbb{U}_{it}^\top),\ \Omega_{iT,\beta\theta} \equiv \frac{1}{T}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}(\mathbb{U}_{is}\mathbb{W}_{it}^\top),\ \text{and } \Omega_{iT,\theta\theta} \equiv \frac{1}{T}\sum_{s=1}^T\sum_{t=1}^T \mathbb{E}(\mathbb{W}_{is}\mathbb{W}_{it}^\top).$$

Define

$$\mathbb{B}_{NT} \equiv \begin{bmatrix} \mathbb{B}_{1NT} \\ \vdots \\ \mathbb{B}_{K^0NT} \\ \mathbb{B}_{\theta NT} \end{bmatrix} = \begin{bmatrix} \mathbb{B}_{1,1NT} - \mathbb{B}_{2,1NT} \\ \vdots \\ \mathbb{B}_{1,K^0NT} - \mathbb{B}_{2,K^0NT} \\ \mathbb{B}_{1,\theta NT} - \mathbb{B}_{2,\theta NT} \end{bmatrix},$$

$$\Omega_{NT} \equiv \begin{bmatrix} \frac{1}{N_1}\sum_{i\in G_1^0}\Omega_{iT,\beta\beta} & \cdots & 0 & \frac{1}{N_1}\sum_{i\in G_1^0}\Omega_{iT,\beta\theta} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{N_{K^0}}\sum_{i\in G_{K^0}^0}\Omega_{iT,\beta\beta} & \frac{1}{N_{K^0}}\sum_{i\in G_{K^0}^0}\Omega_{iT,\beta\theta} \\ \frac{1}{N}\sum_{i\in G_1^0}\Omega_{iT,\beta\theta}^\top & \cdots & \frac{1}{N}\sum_{i\in G_{K^0}^0}\Omega_{iT,\beta\theta}^\top & \frac{1}{N}\sum_{i=1}^N\Omega_{iT,\theta\theta} \end{bmatrix},$$

$$\mathbb{H}_{NT}(\boldsymbol{\beta},\theta) \equiv \begin{bmatrix} \frac{1}{N_1}\sum_{i\in G_1^0}H_{i,\beta\beta}(\beta_i,\theta) & \cdots & 0 & \frac{1}{N_1}\sum_{i\in G_1^0}H_{i,\beta\theta}(\beta_i,\theta) \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{N_{K^0}}\sum_{i\in G_{K^0}^0}H_{i,\beta\beta}(\beta_i,\theta) & \frac{1}{N_{K^0}}\sum_{i\in G_1^0}H_{i,\beta\theta}(\beta_i,\theta) \\ \frac{1}{N}\sum_{i=1}^N H_{i,\theta\beta}(\beta_i,\theta) & \cdots & \frac{1}{N}\sum_{i=1}^N H_{i,\theta\beta}(\beta_i,\theta) & \frac{1}{N}\sum_{i=1}^N H_{i,\theta\theta}(\beta_i,\theta) \end{bmatrix}, \tag{3.2}$$

where

$$\mathbb{B}_{1,kNT} = \left(N_kT^3\right)^{-1/2}\sum_{i\in G_k^0}\sum_{s=1}^T\sum_{t=1}^T \mathbb{U}_{it}^\mu S_{iV}^{-1}V_{is},$$

$$[\mathbb{B}_{2,kNT}]_j = \frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right)^\top S_{iV}^{-1}S_{iU2,j}S_{i,V}^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right) - \frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}S_{iU}S_{iV}^{-1}R_{iV},$$

$$\mathbb{B}_{1,\theta NT} = \left(NT^3\right)^{-1/2}\sum_{i=1}^N\sum_{s=1}^T\sum_{t=1}^T \mathbb{W}_{it}^\mu S_{iV}^{-1}V_{is},$$

$$[\mathbb{B}_{2,\theta NT}]_j = \frac{1}{2\sqrt{NT}}\sum_{i=1}^N\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right)^\top S_{iV}^{-1}S_{iW2,j}S_{i,V}^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it}\right) - \frac{1}{2\sqrt{NT}}\sum_{i=1}^N S_{iW}S_{iV}^{-1}R_{iW},$$

$$H_{i,\beta\beta}(\beta_i,\theta) = \frac{1}{T}\sum_{t=1}^T\left[U^\beta(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta) + U^\mu(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta)\frac{\partial\mu_i(\beta_i,\theta)}{\partial\beta_i^\top}\right],$$

$$H_{i,\beta\theta}(\beta_i,\theta) = \frac{1}{T}\sum_{t=1}^T\left[U^\theta(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta) + U^\mu(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta)\frac{\partial\mu_i(\beta_i,\theta)}{\partial\theta^\top}\right],$$

$$H_{i,\theta\beta}(\beta_i,\theta) = \frac{1}{T}\sum_{t=1}^T\left[W^\beta(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta) + W^\mu(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta)\frac{\partial\mu_i(\beta_i,\theta)}{\partial\beta_i^\top}\right],$$

$$H_{i,\theta\theta}(\beta_i,\theta) = \frac{1}{T}\sum_{t=1}^T\left[W^\theta(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta) + W^\mu(w_{it};\beta_i,\mu_i(\beta_i,\theta),\theta)\frac{\partial\mu_i(\beta_i,\theta)}{\partial\theta^\top}\right].$$

Hereafter, $[A]_j$ denotes the jth element of the vector A, $[R_{iV}]_j = (\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it})^\top S_{iV}^{-1}S_{iV2,j}S_{iV}^{-1}(\frac{1}{\sqrt{T}}\times\sum_{t=1}^T V_{it})$, and $[R_{iW}]_j = (\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it})^\top S_{iV}^{-1}S_{iW2,j}S_{iV}^{-1}(\frac{1}{\sqrt{T}}\sum_{t=1}^T V_{it})$.

As we will see, Ω_{NT} and $\mathbb{H}_{NT}\left(\boldsymbol{\beta}^0, \theta^0\right)$ enter the asymptotic variance of our estimators and $\mathbb{B}_{NT}$ contributes to the asymptotic bias.

To study the asymptotic distribution of our estimators, we add an assumption.

Assumption A4 (i) $\Omega \equiv \lim_{(N,T)\to\infty} \Omega_{NT}$ exists and is positive definite.

(ii) $\mathbb{H} \equiv \lim_{(N,T)\to\infty} \mathbb{E}[\mathbb{H}_{NT}\left(\boldsymbol{\beta}^0, \theta^0\right)]$ exists and is nonsingular.

Assumption A4 is needed to derive the asymptotic bias and variance of the post-classification estimators $\hat{\alpha}_k$'s and $\hat{\theta}$. Define the oracle estimators $\hat{\alpha}_k^*$'s and $\hat{\theta}^*$ of α_k and θ that are obtained with $\hat{K}$ and $\hat{G}_k(\hat{K})$ in (2.10) being replaced by K^0 and G_k^0. The following theorem indicates that these two set of estimators are asymptotically equivalent.

Theorem 3.4 (Asymptotic distribution)**.** *Suppose that Assumptions A1–A4 hold. By using the SBSA 1 in Section 2.2 and the information criteria in (2.8), the final estimators $\hat{\alpha}_k$'s and $\hat{\theta}$ are asymptotically equivalent to the oracle estimators $\hat{\alpha}_k^*$'s and $\hat{\theta}^*$. In particular, conditional on the large-probability event $\{\hat{K} = K^0\}$ we have*

$$D_{NT}\begin{bmatrix} \hat{\alpha}_1 - \alpha_1^0 \\ \vdots \\ \hat{\alpha}_{K^0} - \alpha_{K^0}^0 \\ \hat{\theta} - \theta^0 \end{bmatrix} + \mathbb{H}_{NT}^{-1}\mathbb{B}_{NT} \xrightarrow{D} N\left(0, \mathbb{H}^{-1}\Omega(\mathbb{H}^{-1})^\top\right), \tag{3.3}$$

where $D_{NT} = \text{diag}(\sqrt{N_1 T} I_p, \ldots, \sqrt{N_{K^0} T} I_p, \sqrt{NT} I_q)$ *and* $\mathbb{H}_{NT} = \mathbb{H}_{NT}(\boldsymbol{\beta}^0, \theta^0)$.

Note that we explicitly write elements of $\mathbb{B}_{NT}$ as the difference between two terms that are derived from the first- and second-order Taylor expansion of the profile log-likelihood estimating equation, respectively. Comparing the above results with those in Hahn and Kuersteiner (2011) and SSP, our asymptotic bias and variance formulae are a little bit more complicated than theirs due to the presence of the common parameter θ. In the absence of θ, both formulae can be simplified and one can easily verify that in this case the asymptotic bias and variance of $\hat{\alpha}_k$'s are the same as those of the group-specific parameter estimators in SSP.

To make inference, we need to estimate both the asymptotic bias and variance consistently. Given the fact that the elements of $\mathbb{H}_{NT}$ and $\mathbb{B}_{NT}$ share similar structures as those in SSP, one can follow SSPb and obtain the analytical formulae for both estimates and justify their consistency. Alternatively, we can use the jackknife method to correct bias. See Hahn and Newey (2004) and Dhaene and Jochmans (2015) for static and dynamic panels, respectively.

4 An improved algorithm

In this section we consider an improved algorithm that is based on the spectral decomposition of the $N \times N$ matrix $\tilde{\boldsymbol{D}}_N = N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top$. We first explain why the eigenvectors associated with the few largest eigenvalues of $\tilde{\boldsymbol{D}}_N$ contain the individual's group information. Then we show that we can

apply the SBSA to these eigenvectors to infer the group membership for all individuals w.p.a.1. The post-classification estimation and inference then follows directly from the previous section.

4.1 Spectral decomposition

Define the $K^0 \times K^0$ matrix and $N \times N$ matrix:

$$\boldsymbol{A} \equiv \boldsymbol{\alpha}^0 \boldsymbol{\alpha}^{0\top} = \begin{pmatrix} \alpha_1^{0\top}\alpha_1^0 & \cdots & \alpha_1^{0\top}\alpha_{K^0}^0 \\ \vdots & \ddots & \vdots \\ \alpha_{K^0}^{0\top}\alpha_1^0 & \cdots & \alpha_{K^0}^{0\top}\alpha_{K^0}^0 \end{pmatrix} \text{ and } \boldsymbol{D}_N \equiv N^{-1}\boldsymbol{\beta}^0\boldsymbol{\beta}^{0\top}. \tag{4.1}$$

Define an $N \times K^0$ matrix $\boldsymbol{Z}_N \in \{0,1\}^{N\times K^0}$ that has exactly one 1 in each row and N_k 1's in column k where $k = 1, \ldots, K^0$. Let $z_i^\top$ denote the ith row of $\boldsymbol{Z}_N$ for $i = 1, \ldots, N$. The position of the single 1 in z_i indicates the group membership of individual i. For example, $z_i^\top = (1, 0, \ldots, 0)$ indicates that individual i belongs to Group 1 and $z_i^\top = (0, 0, \ldots, 1)$ indicates that individual i belongs to Group K^0. Apparently, we have

$$\boldsymbol{D}_N = N^{-1}\boldsymbol{Z}_N\boldsymbol{A}\boldsymbol{Z}_N^\top. \tag{4.2}$$

The expression in (4.2) helps us to link the panel structure model with the stochastic block model (SBM) that is widely used for community detection in the network literature**.** In a SBM that contains N nodes (vertices) and K communities (blocks), each node belongs to one of the K communities, and the probability for two nodes to form a link only depends on the community membership. In comparison of the SBM, $\boldsymbol{Z}_N$ stores the individuals' group membership in our model and nodes' community membership in a SBM. The matrix $\boldsymbol{A}$ here is analogous to the probability matrix that contains the probability of edges within and between blocks in a SBM; but we do not restrict elements of $\boldsymbol{A}$ to lie between 0 and 1. In both cases, the main interest is to estimate $\boldsymbol{Z}_N$ based on some sample information.

Various spectral clustering algorithms have been proposed for community detection based on a SBM. It has been suggested that the eigenvectors corresponding to the few largest eigenvalues of certain matrix associated with the adjacency matrix reveal the clusters of interest. For example, Rohe, Chatterjee, and Yu (2011) work on the eigenvectors of a normalized adjacency matrix. This motivates us to consider the eigenvectors of the sample analogue of $\boldsymbol{D}_N$, the counterpart of the adjacent matrix, to identify the latent group structure.

To appreciate the advantages of using eigenvectors to identify the latent group structures, we consider the example below.

Example 4.1 (When $p > K^0$)**.** This is a case when implementing SBSA on the eigenvectors is generally better than on $\tilde{\boldsymbol{\beta}}$. If the difference between different columns of the $p \times K^0$ matrix $\boldsymbol{\alpha}^{0\top}$ is small for each row, then it is difficult to use SBSA 1 to achieve group identification. Nevertheless, the eigenvectors associated with the few largest eigenvalues of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top$ (or $\boldsymbol{D}_N$) summarize all the useful group information and implementing the SBSA on the eigenvectors tend to outperform

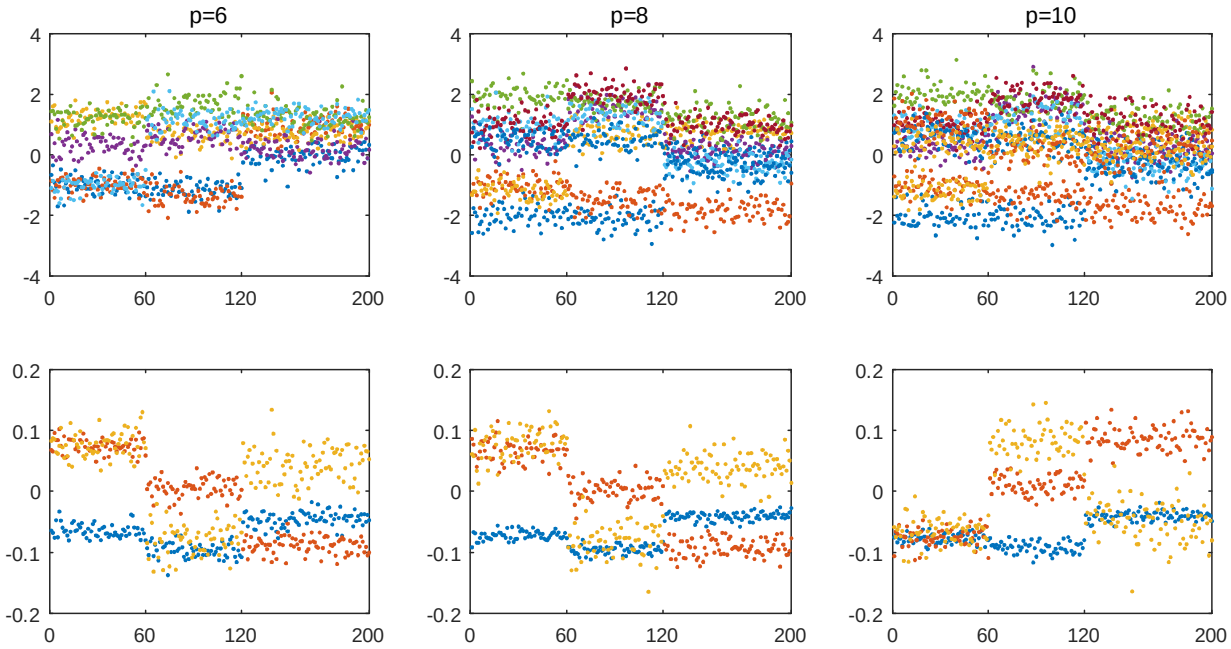

Figure 1: Comparison of the plots of the p columns in the preliminary estimates $\tilde{\boldsymbol{\beta}}$ with the three eigenvectors of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^{\top}$ associated with its three largest eigenvalues when $p > K^0$: row 1 for preliminary estimates and row 2 for eigenvectors

that based on the original $\tilde{\boldsymbol{\beta}}$ matrix. Due to limited space, we only consider $N = 200$ and $T = 20$ for a linear DGP with three groups ($K^0 = 3$) and p regressors, where the group ratio is $3 : 3 : 4$. We consider three values of p: $6, 8, 10$. In Figure 1, the first row plots different columns in $\tilde{\boldsymbol{\beta}}$ for $p = 6, 8, 10$, and the second row plots the three eigenvectors corresponding to the three largest eigenvalues of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^{\top}$ for each p. The true group coefficients are not displayed here to save space. From the figure, we can tell that the eigenvectors reveal the true group information much more clearly than $\tilde{\boldsymbol{\beta}}$. This is especially true when p is large (say $p = 10$).

Let K^* denote the number of strictly positive eigenvalues of $\boldsymbol{A}$. Apparently, $K^* \leq \min\left(K^0, p\right)$. We consider the spectral decomposition of $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{u}\Lambda\boldsymbol{u}^{\top},$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_{K^*})$ is a $K^* \times K^*$ matrix that contains the nonzero eigenvalues of $\boldsymbol{A}$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{K^*} > 0$, and the columns of $\boldsymbol{u}$ contain the eigenvectors of $\boldsymbol{A}$ such that $\boldsymbol{u}^{\top}\boldsymbol{u} = I_{K^*}$. Interestingly, Assumption A2(i), $\min_{1\leq k<k'\leq K^0} \|\alpha_k^0 - \alpha_{k'}^0\| > c_L > 0$, ensures that the K^0 rows of $\boldsymbol{u}$ are distinct from each other. See the proof of Lemma 4.1 below. Similarly, we consider the spectral decomposition of $\boldsymbol{D}_N$

$$\boldsymbol{D}_N = N^{-1}\boldsymbol{U}_N\Sigma_N\boldsymbol{U}_N^{\top} = N^{-1}\boldsymbol{U}_{1,N}\Sigma_{1,N}\boldsymbol{U}_{1,N}^{\top},$$

where $\Sigma_N = \text{diag}(\mu_{1N}, \ldots, \mu_{K^*N}, 0, \ldots, 0)$ is a $p \times p$ matrix that contains the eigenvalues of $\boldsymbol{D}_N$ in descending order along its diagonal, $\Sigma_{1,N} = \text{diag}(\mu_{1N}, \ldots, \mu_{K^*N})$, the columns of $\boldsymbol{U}_N$ contain the eigenvectors of $\boldsymbol{D}_N$ associated with the eigenvalues in Σ_N, $\boldsymbol{U}_N = (\boldsymbol{U}_{1,N}, \boldsymbol{U}_{2,N})$, and $\boldsymbol{U}_N^\top \boldsymbol{U}_N = I_p$. The following lemma establishes the link between the eigenvalues and eigenvectors of $\boldsymbol{A}$ and those of $\boldsymbol{D}_N$.

Lemma 4.1. *Let $\boldsymbol{A}$, $\boldsymbol{D}_N$, Λ, $\Sigma_{1,N}$, $\boldsymbol{u}$ and $\boldsymbol{U}_{1,N}$ be defined as above. Then there exists a nonsingular matrix $\boldsymbol{S} \equiv \boldsymbol{S}_N$ such that (i) the diagonal matrix $\Sigma_{1,N}$ can be written as $\boldsymbol{S}^{-1}\Lambda(\boldsymbol{S}^{-1})^\top$, (ii) $\boldsymbol{U}_{1,N} = N^{-1/2}\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S}$, (iii) $\boldsymbol{S}$ is given by $(N^{-1/2}\boldsymbol{U}_{1,N}^\top \boldsymbol{Z}_N\boldsymbol{u})^{-1}$, and (iv) $z_i^\top \boldsymbol{u}\boldsymbol{S} = z_j^\top \boldsymbol{u}\boldsymbol{S}$ if and only $z_i = z_j$ for $i, j = 1, 2, \ldots, N$.*

The last result in Lemma 4.1 is obvious if $\boldsymbol{u}\boldsymbol{S}$ is a nonsingular square matrix. In this case, there exists a one-to-one map between $\boldsymbol{U}_{1,N}$ and $\boldsymbol{Z}_N$. In the general case, we allow $K^* < K^0$ so that $\boldsymbol{u}\boldsymbol{S}$ has rank K^* only, and we show in the proof of the above lemma that the rows of $\boldsymbol{u}\boldsymbol{S}$ are distinct from each other. This ensures that the rows of $\boldsymbol{U}_{1,N}$ contain the same group information as $\boldsymbol{Z}_N$. Therefore, we can infer each individual's group membership based on the eigenvector matrix $\boldsymbol{U}_{1,N}$ if $\boldsymbol{D}_N$ is observed.

In practice, $\boldsymbol{D}_N$ is not observed. But we can estimate it by

$$\tilde{\boldsymbol{D}}_N \equiv N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top.$$

Consider the spectral decomposition of $\tilde{\boldsymbol{D}}_N$: $\tilde{\boldsymbol{D}}_N = \tilde{\boldsymbol{U}}_N\tilde{\Sigma}_N\tilde{\boldsymbol{U}}_N^\top$, where $\tilde{\Sigma}_N = \text{diag}(\tilde{\mu}_{1,N}, \ldots, \tilde{\mu}_{p,N})$ contains the first p eigenvalues of $\tilde{\boldsymbol{D}}_N$ in descending order. By Theorem 3.1, we can readily show that $\|\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N\| = O_P(T^{-1/2})$, ensuring that $\max_{1\le\ell\le N}|\tilde{\mu}_{\ell,N} - \mu_{\ell,N}| \le \|\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N\| = O_P\left(T^{-1/2}\right)$, where $\tilde{\mu}_{\ell,N}$ and $\mu_{\ell,N}$ denote the ℓth largest eigenvalues of $\tilde{\boldsymbol{D}}_N$ and $\boldsymbol{D}_N$, respectively. To take into account the possibility of estimating a zero eigenvalue of $\boldsymbol{D}_N$ by a positive value, we have to ensure that μ_{K^*N} is not too close to zero in order to identify the nonzero eigenvalues of $\boldsymbol{D}_N$ and apply the Davis-Kahan theorem (see, e.g., the sin θ theorem in Davis and Kahan (1970), Chapter VII in Bhatia (1997), Proposition 2.1 in Rohe, Chatterjee, and Yu (2011), Theorem 3 in Yu, Wang, and Samworth (2015)).

Recall $\lambda_j(A)$ denotes the jth largest eigenvalue of a symmetric matrix A. For clarity, we continue to assume that K^0 is fixed. In this case, it is natural to assume that $\lambda_{K^*}(\boldsymbol{A}) = \lambda_{K^*}\left(\boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\top}\right) \ge c$ for some constant $c > 0$. Noting that AB and BA share the same set of nonzero eigenvalues, we have

$$\begin{aligned}\mu_{K^*N} = \lambda_{K^*}(\boldsymbol{D}_N) &= \lambda_{K^*}\left(N^{-1}\boldsymbol{Z}_N\boldsymbol{A}\boldsymbol{Z}_N^\top\right) = \lambda_{K^*}\left(\boldsymbol{A}N^{-1}\boldsymbol{Z}_N^\top\boldsymbol{Z}_N\right) \\ &\ge \lambda_{K^*}(\boldsymbol{A})\,\lambda_{\min}\left(N^{-1}\boldsymbol{Z}_N^\top\boldsymbol{Z}_N\right) \ge c \min_{1\le k\le K^0} N_k/N. \end{aligned} \tag{4.3}$$

It follows that $\lim_{N\to\infty}\mu_{K^*N} \ge c\min_{1\le k\le K^0}\tau_k > 0$ under Assumption A2(ii). Since only the eigenvectors that are associated with the K^* nonzero eigenvalues of $\boldsymbol{D}_N$ can contain the group information, we will restrict our attention to the eigenvectors associated with the first $\mathcal{K}_N$ eigenvalues of $\tilde{\boldsymbol{D}}_N$ such that $\lambda_{\mathcal{K}_N}(\tilde{\boldsymbol{D}}_N) \ge c_N$, where c_N is a positive sequence that converges to zero at

a slow rate, e.g., $c_N = 0.1/\log N$. By choosing such a tuning parameter, we can effectively avoid using eigenvectors associated with the eigenvalues of $\tilde{\boldsymbol{D}}_N$ whose population values are zero. To see this, notice that when $\mathcal{K}_N > K^*$, $\lambda_{\mathcal{K}_N}(\tilde{\boldsymbol{D}}_N)$ converges to zero in probability at rate $T^{-1/2}$. So it is easy to show that $\mathcal{K}_N = K^*$ w.p.a.1.

Given $\mathcal{K}_N$, we decompose $\tilde{\boldsymbol{U}}_N$ and $\tilde{\Sigma}_N$ as follows: $\tilde{\boldsymbol{U}}_N = (\tilde{\boldsymbol{U}}_{1,N}, \tilde{\boldsymbol{U}}_{2,N})$ and $\tilde{\Sigma}_N = \text{diag}(\tilde{\Sigma}_{1,N}, \tilde{\Sigma}_{2,N})$, where $\tilde{\boldsymbol{U}}_{1,N}$ is an $N \times \mathcal{K}_N$ matrix and $\tilde{\Sigma}_{1,N}$ contains the largest $\mathcal{K}_N$ eigenvalues of $\tilde{\boldsymbol{D}}_N$ along its diagonal in descending order. Let $\tilde{u}_i^\top = (\tilde{u}_{1,i}^\top, \tilde{u}_{2,i}^\top)$ and $u_i^\top = (u_{1,i}^\top, u_{2,i}^\top)$ denote the ith row of $\tilde{\boldsymbol{U}}_N = (\tilde{\boldsymbol{U}}_{1,N}, \tilde{\boldsymbol{U}}_{2,N})$ and $\boldsymbol{U}_N = (\boldsymbol{U}_{1,N}, \boldsymbol{U}_{2,N})$, respectively.

To state the next theorem, we add the following assumption.

Assumption A5 There exist a positive constant c such that $\lambda_{K^*}(\boldsymbol{A}) \geq c$.

The main result in this subsection is summarized in the following theorem.

Theorem 4.2. *Suppose that Assumptions A1–A5 hold. Then $\mathcal{K}_N = K^*$ w.p.a.1. Furthermore, conditional on $\mathcal{K}_N = K^*$, there exists a sequence of $K^* \times K^*$ orthogonal matrices O_N such that* $\max_{1\leq i\leq N} \sqrt{N}\,\|\tilde{u}_{1,i} - O_N u_{1,i}\| = O_P\left(T^{-1/2}(\ln T)^3\right)$.

An immediate implication of Theorem 4.2 is $\|\tilde{\boldsymbol{U}}_{1,N} - \boldsymbol{U}_{1,N}O_N\| = O_P\left(T^{-1/2}(\ln T)^3\right) = o_P(1)$, and like $\boldsymbol{U}_{1,N}$, $\tilde{\boldsymbol{U}}_{1,N}$ contains the true group information for all individuals. As a result, we can consider the SBSA based on $\tilde{\boldsymbol{U}}_{1,N}$ instead of $\tilde{\boldsymbol{\beta}}$.

4.2 An eigenvector-based SBSA

Since $\tilde{\boldsymbol{U}}_{1,N}$ contains the group membership for all individuals, we implement the SBSA based on it. Let $\tilde{\boldsymbol{U}}_{1,N} = (\tilde{\boldsymbol{U}}_{\cdot 1}, \ldots, \tilde{\boldsymbol{U}}_{\cdot \mathcal{K}_N})$ and $\boldsymbol{U}_{1,N} = (\boldsymbol{U}_{\cdot 1}, \ldots, \boldsymbol{U}_{\cdot \mathcal{K}_N})$.[2] Let U_{ij} and $\tilde{U}_{ij}$ denote the ith element of $\boldsymbol{U}_{\cdot j}$ and $\tilde{\boldsymbol{U}}_{\cdot j}$, respectively. We sort the N elements of $\tilde{\boldsymbol{U}}_{\cdot j}$ in ascending order and denote the order statistics by

$$\tilde{U}_{\pi_j(1),j} \leq \tilde{U}_{\pi_j(2),j} \leq \cdots \leq \tilde{U}_{\pi_j(N),j}, \tag{4.4}$$

where $\{\pi_j(1), \ldots, \pi_j(N)\}$ is a permutation of $\{1, \ldots, N\}$ that is implicitly determined by the order relation in (4.4). Let

$$\tilde{\mathcal{S}}_{i,l}(j) \equiv \{\tilde{U}_{\pi_j(i),j}, \tilde{U}_{\pi_j(i+1),j}, \ldots, \tilde{U}_{\pi_j(l),j}\}$$

where $1 \leq i < l \leq N$.

Let

$$\bar{U}_{i,l}(j) = \frac{1}{l-i+1}\sum_{i'=i}^{l} \tilde{U}_{\pi_j(i'),j} \quad \text{and } \tilde{V}_{i,l}(j) \equiv \frac{1}{l-i}\sum_{i'=i}^{l}[\tilde{U}_{\pi_j(i'),j} - \bar{\mathcal{U}}_{i,l}(j)]^2$$

denote the sample mean and variance of the subsample $\tilde{\mathcal{S}}_{i,l}(j)$. Define

$$\tilde{S}_{i,l}(j,m) = \frac{1}{l-i+1}\left\{\sum_{i'=i}^{m}\left[\tilde{U}_{\pi_j(i'),j} - \bar{U}_{i,m}(j)\right]^2 + \sum_{i'=m+1}^{l}\left[\tilde{U}_{\pi_j(i'),j} - \bar{U}_{m+1,l}(j)\right]^2\right\}. \tag{4.5}$$

[2] To account for the scale effect, we use $\tilde{\boldsymbol{\beta}}' = (\tilde{\boldsymbol{\beta}}'_{\cdot 1}, \ldots, \tilde{\boldsymbol{\beta}}'_{\cdot p})$ where $\tilde{\boldsymbol{\beta}}'_{\cdot j} = \tilde{\boldsymbol{\beta}}_{\cdot j}/\sqrt{\bar{\sigma}^2_{1,N}(j)}$, $j = 1, \ldots, p$, instead of $\tilde{\boldsymbol{\beta}}$ in calculating the eigenvectors $\tilde{\boldsymbol{U}}_{1,N}$. Recall that $\bar{\sigma}^2_{1,N}(j)$ is defined in Section 2.2.

We propose to adopt the following eigenvector-based SBSA to estimate $\mathcal{G}^0$.

Sequential Binary Segmentation Algorithm 2 (SBSA 2)

1. Let $K \in [1, K_{\max}]$. When $K = 1$, there is only one group with the estimate $\tilde{G}_1(1) = \{1, \ldots, N\}$.

2. When $K = 2$, let $\tilde{j}_1 = \arg\max_{1 \le j \le \mathcal{K}_N} \tilde{V}_{1,N}(j)$. Given $\tilde{j}_1$, we solve the following minimization problem
$$\tilde{m}_1 \equiv \underset{1 \le m < N}{\arg\min}\, \tilde{S}_{1,N}(\tilde{j}_1, m).$$
Now we have two segments – $\tilde{G}_1(2) = \tilde{\mathcal{S}}_{1,\tilde{m}_1}(\tilde{j}_1)$ and $\tilde{G}_2(2) = \tilde{\mathcal{S}}_{\tilde{m}_1+1,N}(\tilde{j}_1)$.

3. When $K \ge 3$, we use $\tilde{m}_1, \ldots, \tilde{m}_{K-2}$ denote the break points detected in the previous steps such that $\tilde{m}_1 < \cdots < \tilde{m}_{K-2}$ (perhaps after relabeling) the $K-2$ break points that have been detected so far. Define
$$\begin{aligned}
\tilde{j}_{K-1} &\equiv \underset{1 \le j \le \tilde{\mathcal{K}}_N}{\arg\max} \sum_{k=1}^{K-1} \tilde{V}_{\tilde{m}_{k-1}+1,\tilde{m}_k}(j), \\
\tilde{m}_{K-1}(k) &\equiv \underset{\tilde{m}_{k-1}+1 \le m < \tilde{m}_k}{\arg\min} \tilde{S}_{\tilde{m}_{k-1}+1,\tilde{m}_k}(\tilde{j}_{K-1}, m) \text{ for } k = 1, \ldots, K-1,
\end{aligned}$$
where $\tilde{m}_0 = 0$, $\tilde{m}_{K-1} = N$, and we suppress the dependence of $\tilde{m}_{K-1}(k)$ on $\tilde{j}_{K-1}$. Then $\tilde{m}_{K-1}(k)$ divides $\tilde{G}_k(K-1)$ into two subsegments, which are labeled as $\tilde{G}_{k1}(K-1)$ and $\tilde{G}_{k2}(K-1)$ respectively. Calculate for $k = 1, \ldots, K-1$,
$$\begin{aligned}
\tilde{S}_{K-1}(k) \equiv & \sum_{i \in \tilde{G}_{k1}(K-1)} \left[\tilde{U}_{i,\tilde{j}_{K-1}} - \bar{U}_{\tilde{G}_{k1}(K-1)}(\tilde{j}_{K-1})\right]^2 \\
& + \sum_{i \in \tilde{G}_{k2}(K-1)} \left[\tilde{U}_{i,\tilde{j}_{K-1}} - \bar{U}_{\tilde{G}_{k2}(K-1)}(\tilde{j}_{K-1})\right]^2 \\
& + \sum_{1 \le l \le K-1, l \ne k} \sum_{i \in \tilde{G}_l(K-1)} \left[\tilde{U}_{i,\tilde{j}_{K-1}} - \bar{U}_{\tilde{G}_l(K-1)}(\tilde{j}_{K-1})\right]^2,
\end{aligned}$$
where, e.g., $\bar{U}_{\tilde{G}_{k1}(K-1)}(\tilde{j}_{K-1}) = |\tilde{G}_{k1}(K-1)|^{-1} \sum_{i \in \tilde{G}_{k1}(K-1)} \tilde{U}_{i,\tilde{j}_{K-1}}$. Let
$$\tilde{k} = \underset{1 \le k \le K-1}{\arg\min}\, \tilde{S}_{K-1}(k).$$
We now obtain the $K-1$ break points and the K segments given by $\{\tilde{m}_1, \ldots, \tilde{m}_{K-2}, \tilde{m}_{K-1}(\tilde{k})\}$ and $\{\tilde{G}_1(K-1), \ldots, \tilde{G}_{\tilde{k}-1}(K-1), \tilde{G}_{\tilde{k}1}(K-1), \tilde{G}_{\tilde{k}2}(K-1), \tilde{G}_{\tilde{k}+1}(K-1), \ldots, \tilde{G}_{K-1}(K)\}$, respectively. Relabel these $K-1$ break points as $\{\tilde{m}_1, \ldots, \tilde{m}_{K-1}\}$ such that $\tilde{m}_1 < \tilde{m}_2 < \cdots < \tilde{m}_{K-1}$, and the corresponding K groups as $\{\tilde{G}_1(K), \tilde{G}_2(K), \ldots, \tilde{G}_K(K)\}$.

4. Repeat the last step until $K = K^{\max}$.

Of course if K^0 is known *a priori*, we can set $K^{\max} = K^0$. At the end of the SBSA, we obtain the $\hat{\mathcal{G}}\left(K^0\right) \equiv \{\tilde{G}_1, \tilde{G}_2, \ldots, \tilde{G}_{K^0}\}$ as the estimates of the true group structure $\mathcal{G}^0$. Otherwise, we can estimate K^0 either based on SBSA 1 or SBSA 2.

Let $\hat{\boldsymbol{\beta}}^*(K), \hat{\boldsymbol{\mu}}^*(K)$, and $\hat{\theta}^*(K)$ be defined analogously to $\hat{\boldsymbol{\beta}}(K), \hat{\boldsymbol{\mu}}(K)$, and $\hat{\theta}(K)$, now with the estimated group based on SBSA 2. We can estimate K^0 by minimizing the following BIC-type information criterion

$$\mathrm{IC}_2(K) = 2L_{NT}(\hat{\boldsymbol{\beta}}^*(K), \hat{\boldsymbol{\mu}}^*(K), \hat{\theta}^*(K)) + pK \cdot \rho_{NT}. \tag{4.6}$$

Let

$$\tilde{K} \equiv \underset{1 \le K \le K^{\max}}{\arg\min}\ \mathrm{IC}_2(K) \text{ and } \tilde{\mathcal{G}} \equiv \tilde{\mathcal{G}}(\tilde{K}) \equiv \{\tilde{G}_1(\tilde{K}), \tilde{G}_2(\tilde{K}), \ldots, \tilde{G}_{\tilde{K}}(\tilde{K})\}. \tag{4.7}$$

We will show that $P(\tilde{K} = K^0) \to 1$ and $P(\tilde{\mathcal{G}} = \mathcal{G}^0)$ as $(N, T) \to \infty$.

Given $\tilde{K}$ and $\tilde{\mathcal{G}}$, we consider the constrained minimization problem in (2.7) with K being replaced by $\tilde{K}$ and obtain the final estimate of $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, θ, and $\boldsymbol{\alpha}$. In particular, we denote the estimates as $\boldsymbol{\alpha}$ and θ as $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\theta}$, which can be obtained as the minimizer of (2.10) with $\hat{K}$ and $\hat{G}_k(\hat{K})$ being replaced by $\tilde{K}$ and $\tilde{G}_k(\tilde{K})$. Let $\tilde{\alpha}_k$ denote the kth column of $\tilde{\boldsymbol{\alpha}}^\top$. The following section reports the asymptotic properties of $\tilde{\mathcal{G}}(K^0)$, $\tilde{K}$ and $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\theta}$.

4.3 Asymptotic properties

In this subsection, we first state Theorems 4.3–4.5 which parallel Theorems 3.2–3.4 in Section 3, and then provide some intuitive explanations on why they hold.

Theorem 4.3 (Classification consistency)**.** *Suppose Assumptions A1–A2 and A5 hold. Suppose the true number of groups is known to be K^0. Let $\tilde{\mathcal{G}}(K^0) = \{\tilde{G}_1(K^0), \ldots, \tilde{G}_{K^0}(K^0)\}$ be the estimated group structure based on the SBSA 2. Then $P(\tilde{\mathcal{G}}(K^0) = \mathcal{G}^0) \to 1$ as $(N, T) \to \infty$.*

Theorem 4.4 (Consistency of the information criterion)**.** *Suppose Assumptions A1–A3 and A5 hold. Let $\tilde{K}$ be as defined in (4.7). Then $P(\tilde{K} = K^0) \to 1$ as $(N, T) \to \infty$.*

Theorem 4.5 (Asymptotic distribution)**.** *Suppose that Assumptions A1–A5 hold. By using the SBSA 2 in Section 4.2 and the information criteria in (4.6), the final estimators $\tilde{\alpha}_k$'s and $\tilde{\theta}$ are asymptotically equivalent to the oracle estimators $\hat{\alpha}_k^*$'s and $\hat{\theta}^*$. In particular, conditional on the large-probability event $\{\tilde{K} = K^0\}$, the asymptotic distribution of $D_{NT}((\tilde{\alpha}_1 - \alpha_1^0)^\top, \ldots, (\tilde{\alpha}_{K^0} - \alpha_{K^0}^0)^\top, (\tilde{\theta} - \theta^0)^\top)^\top$ is identical to $D_{NT}((\hat{\alpha}_1 - \alpha_1^0)^\top, \ldots, (\hat{\alpha}_{K^0} - \alpha_{K^0}^0)^\top, (\hat{\theta} - \theta^0)^\top)^\top$ studied in Theorem 3.4.*

Combining the results in Theorems 4.3–4.4, we can recover the true group structure $\mathcal{G}^0$ w.p.a.1 by using the SBSA 2 and IC_2 defined in (4.6). From the proof of Theorem 3.2, we can tell that the key condition to ensure the consistency of classification is the uniform consistency of the preliminary estimates $\tilde{\beta}_i$ and the consistency rate does not play a role here. Theorem 4.2

ensures that $\tilde{\boldsymbol{U}}_{1,N}$ contains all the individual's group information that are required and it implies the uniform convergent of $\sqrt{N}(\tilde{u}_{1,i} - Ou_{1,i})$ to zero where O is the probability limit of O_N. This is all that we need in order to infer the individuals group membership consistently. Given the consistency of $\tilde{\mathcal{G}} = \tilde{\mathcal{G}}(\tilde{\mathcal{K}})$ with $\mathcal{G}^0$, the results in Theorem 4.5 can be derived in the same way as those in Theorem 3.4.

5 Monte Carlo simulations

In this section, we evaluate the finite sample performance of our SBSA through simulations.

5.1 Data generating processes

We consider four data generating processes (DGPs) here. DGPs 1–2 specify a linear panel data model while DGPs 3–4 consider a double-censored static panel data model and a left-censored dynamic panel data model, respectively. In all DGPs, the candidate number of individuals are $N = 100,\ 200$ and the time spans are $T = 10,\ 20,\ 40$. We will evaluate all 6 combinations of N and T. The true number of groups is 3, and the group member proportion is given by $|G_1^0| : |G_2^0| : |G_3^0| = 4 : 3 : 3$ in all DGPs.

DGP 1 (Linear panel)**.** The data are generated as

$$y_{it} = x_{it}^\top \beta_i + \mu_i + \varepsilon_{it},$$

where $x_{it} = (x_{1,it}, x_{2,it})^\top$, $x_{1,it} = 0.2\mu_i + e_{1,it}$, $x_{2,it} = 0.2\mu_i + e_{2,it}$, and $e_{1,it}$, $e_{2,it}$, ε_{it} and the fixed effect μ_i are all i.i.d. standard normal and mutually independent of each other. The true coefficients β_i can be classified into 3 groups with true group-specific parameter values given by

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{bmatrix} 0.5 \\ -1 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}, \begin{bmatrix} 0.5 \\ 2 \end{bmatrix} \right).$$

Note that here $\alpha_{1,1}^0 = \alpha_{2,1}^0 = \alpha_{3,1}^0$ but we do not assume that they are known to be common. We want to use this DGP to show our method is robust to this kind of specifications.

DGP 2 (Linear panel with $p = 10$)**.** The data are generated as

$$y_{it} = x_{it}^\top \beta_i + \mu_i + \varepsilon_{it},$$

where x_{it} is a 10×1 vector with the jth element given by $x_{j,it} = 0.2\mu_i + e_{j,it}$, $j = 1, \ldots, 10$, and $e_{j,it}$, ε_{it}, and the fixed effect μ_i are all i.i.d. standard normal and mutually independent of each other. The true coefficients β_i can be classified into 3 groups with true group-specific parameter values given by $\alpha_1^0 = (-1, -1.1, -1.2, 0.3, 2, 1, 0.9, 0.1, 0.1, -0.1)^\top$, $\alpha_2^0 = (-1.1, 0.4, 0.7, 0.6, 1.7, 1.3, 2, 0.5, 0.1, -0.1)^\top$, and $\alpha_3^0 = (0, 1.8, 0.8, 0.2, 1.2, -0.3, 1.9, -0.2, 0.1, -0.1)^\top$. We want to use this DGP to show our SBSA 2 is well suited for the large p case.

DGP 3 (Double-censored static panel)**.** The data are generated according to

$$y_{it} = \text{mami}\left(0, x_{it}^{\top}\beta_i + \mu_i + \varepsilon_{it}, 4\right),$$

where $x_{it} = (x_{1,it}, x_{2,it})^{\top} = (e_{1,it} + 0.1\mu_i, e_{2,it} + 0.1\mu_i)^{\top}$, and $e_{1,it}$, $e_{2,it}$, ε_{it}, μ_i are all independently drawn from the standard normal distribution and are mutually independent of each other. The censored ratio is around 51% (with left censored ratio 50% and right censored ratio 1%). The true group-specific parameter values are

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{bmatrix} 1.5 \\ -1.5 \end{bmatrix}, \begin{bmatrix} -0.5 \\ 0.5 \end{bmatrix}, \begin{bmatrix} -1.8 \\ 1.8 \end{bmatrix}\right).$$

The variance $\sigma^2 = \text{Var}(\varepsilon_{it})$ is modeled as the common parameter across all individuals.

DGP 4 (Dynamic one-side censored panel)**.** The model is

$$y_{it} = \max\left(0, \rho y_{i,t-1} + x_{it}^{\top}\beta_i + \mu_i + \varepsilon_{it}\right),$$

where x_{it}, μ_i, and ε_{it} are generated as in DGP 3. To generate T periods of observations for individual i, we first generate $T + 100$ observations with initial value $y_{i0} = 0$, and then take the last T periods of observations. We discard those individuals which have constant regressor or constant regressand across all T periods. The censored ratio is around 40%. For the parameters, $\rho^0 = 0.4$ and the true group-specific parameter values are

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{bmatrix} -1.2 \\ 1.6 \end{bmatrix}, \begin{bmatrix} 0.6 \\ -0.8 \end{bmatrix}, \begin{bmatrix} 1.5 \\ -1.9 \end{bmatrix}\right).$$

As in DGP 3, σ^2 is modeled as the common parameter across all individuals but we do not assume ρ is common in the estimation procedure.

In all DGPs, we use the information criteria in (2.8) to choose the number of groups. For DGPs 1–2, the information criterion is

$$\text{IC}_1(K) = \sigma^2_{\hat{\mathcal{G}}(K)} + pK\rho_1(NT),$$

where $\rho_1(NT) = \frac{1}{30}\ln(NT)/(NT)^{1/3}$, $\hat{\mathcal{G}}(K) = \{\hat{G}_1(K), \ldots, \hat{G}_K(K)\}$, $\sigma^2_{\hat{\mathcal{G}}(K)} = \frac{1}{NT}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k(K)}$ $\sum_{t=1}^{T}[\tilde{y}_{it} - \tilde{x}_{it}^{\top}\hat{\alpha}_k(K)]^2$, $\tilde{y}_{it} = y_{it} - T^{-1}\sum_{t=1}^{T} y_{it}$, and similarly for $\tilde{x}_{it}$. For DGPs 3–4, the information criterion is

$$\text{IC}_2(K) = 2L_{NT}(\hat{\boldsymbol{\beta}}(K), \hat{\boldsymbol{\mu}}(K), \hat{\theta}(K)) + pK\rho_2(NT), \tag{5.1}$$

where $L_{NT}(\cdot)$ is explained in Section 2, and $\rho_2(NT) = \frac{1}{60}\ln(NT)/(NT)^{1/3}$.

5.2 Simulation results

For all DGPs, results reported here are based on 200 repetitions.

Tables 1 and 2 report the frequency for the selected number of groups based on our information criteria by setting $K^{\max} = 5$. The true number of groups is given by $K^0 = 3$. We compare 4 algorithms: K-means on $\tilde{\boldsymbol{\beta}}$, K-means on the eigenvectors of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top$, SBSA 1 and SBSA 2 for all DGPs. For DGPs 1–2 we also consider C-Lasso.[3] From Tables 1 and 2, we see that for all algorithms, given N, the frequency of choosing the right number of groups increases as T grows. Our methods, especially SBSA 2, enable us to identify the true number of groups with large probability. In DGP 1, SBSA 2 slightly outperforms the C-Lasso and in DGP 2, the opposite is true. Both of them outperform other algorithms significantly. We also see one special property of the binary segmentation algorithm: for fixed T, the frequency of choosing the correct number of groups also increases with N, which is not observed when either the K-means algorithms or SSP's C-Lasso method is employed. In all DGPs under investigation, our information criterion works well for T as small as 10 and it works almost perfectly when $T \geq 20$. In short, our information criterion is quite effective in determining the number of groups.

Suppose the true number of groups K^0 is identified. Now we examine the performance of classification and the post-classification estimators. We follow SSP to define the evaluation criteria. First, we define the percentage of correct classification as $N^{-1}\sum_{k=1}^{K^0}\sum_{i\in\hat{G}_k}\mathbf{1}\{\beta_i^0 = \alpha_k^0\}$, which denotes the percentage of individuals falling into the right group. We show its average value across all replications in columns 4 and 8 of Tables 3 and 4. Columns 5–7 and 9–11 report the performance of the estimates of $\boldsymbol{\alpha}_{\cdot 2}^0 \equiv (\alpha_{1,2}^0, \ldots, \alpha_{K^0,2}^0)^\top$, i.e., the second regressor's coefficient of all groups. We evaluate the performance through three criteria: the root mean squared error (RMSE), bias, and coverage ratio. The RMSE is defined as the weighted average RMSEs of $\alpha_{k,2}^0$, $k = 1, \ldots, K^0$, with weight N_k/N. Specifically, it is $\sum_{k=1}^{K^0}\frac{N_k}{N}\text{RMSE}(\alpha_{k,2}^0)$. Similarly, we define weighted versions of bias, and coverage ratio of the 95% confidence interval estimators. Tables 3 and 4 contain the classification and post-classification results where the oracle estimates are obtained by using the true group structure and the other estimates are obtained based as the post-classification ones.

We summarize some important findings from Tables 3 and 4. First, the percentage of correct classification increases with T for all classification methods under consideration. In particular, for all models under investigation we can achieve almost perfect classification when T increases to 40 by using the improved SBSA 2 method. Second, as expected, the oracle estimates usually have smaller RMSE than the post-classification estimates. Third, like the C-Lasso method, our SBSA 2 method typically outperforms the other methods. As T increases, the RMSEs of the post-classification estimates based on both the C-Lasso method and our SBSA 2 method decrease rapidly and can match those of the oracle ones when $T = 40$. Fourth, the coverage ratios for the post-classification estimates of SBSA 2 improve quickly and get closer to those of the oracle ones

[3] Even in the linear case, the computing time of C-Lasso is around 100 times longer than that of the SBSA methods.

Table 1: The frequency of selecting $K = 1, \ldots, 5$ groups when $K^0 = 3$ and $K^{\max} = 5$

			DGP 1					DGP 2				
	N	T	1	2	**3**	4	5	1	2	**3**	4	5
K-means on $\tilde{\beta}$	100	10	0	0	0.875	0.125	0					
	100	20	0	0	0.845	0.155	0	0	0	0.935	0.060	0.005
	100	40	0	0	0.910	0.090	0	0	0	0.920	0.080	0
	200	10	0	0	0.870	0.120	0.010					
	200	20	0	0	0.855	0.145	0	0	0	0.950	0.050	0
	200	40	0	0	0.865	0.130	0.005	0	0	0.925	0.075	0
K-means on	100	10	0	0.280	0.225	0.205	0.290					
eigenvectors	100	20	0	0.050	0.420	0.360	0.170	0	0	0.975	0.025	0
	100	40	0	0	0.800	0.180	0.020	0	0	0.975	0.025	0
	200	10	0	0.150	0.225	0.270	0.355					
	200	20	0	0.025	0.310	0.370	0.295	0	0	0.985	0.015	0
	200	40	0	0	0.725	0.270	0.005	0	0	0.990	0.010	0
C-Lasso	100	10	0	0	0.995	0.005	0					
	100	20	0	0	1	0	0	0	0	1	0	0
	100	40	0	0	1	0	0	0	0	1	0	0
	200	10	0	0	0.995	0.005	0					
	200	20	0	0	1	0	0	0	0	1	0	0
	200	40	0	0	1	0	0	0	0	1	0	0
SBSA 1	100	10	0	0.010	0.990	0	0					
	100	20	0	0	1	0	0	0	0	0.100	0.890	0
	100	40	0	0	1	0	0	0	0	0.030	0.965	0.005
	200	10	0	0	1	0	0					
	200	20	0	0	1	0	0	0	0	0.065	0.935	0
	200	40	0	0	1	0	0	0	0	0.005	0.995	0
SBSA 2	100	10	0	0	0.995	0.005	0					
	100	20	0	0	1	0	0	0	0	0.990	0.010	0
	100	40	0	0	1	0	0	0	0	1	0	0
	200	10	0	0	1	0	0					
	200	20	0	0	1	0	0	0	0	1	0	0
	200	40	0	0	1	0	0	0	0	1	0	0

Table 2: The frequency of selecting $K = 1, \ldots, 5$ groups when $K^0 = 3$ and $K^{\max} = 5$

			DGP 3					DGP 4				
	N	T	1	2	**3**	4	5	1	2	**3**	4	5
K-means on $\tilde{\beta}$	100	10	0	0.04	0.905	0.055	0	0	0.085	0.7	0.215	0
	100	20	0	0.03	0.875	0.095	0	0	0.06	0.685	0.255	0
	100	40	0	0.005	0.925	0.07	0	0	0.035	0.635	0.32	0.01
	200	10	0	0	0.955	0.045	0	0	0.14	0.705	0.155	0
	200	20	0	0	0.9	0.1	0	0	0.13	0.695	0.175	0
	200	40	0	0	0.905	0.095	0	0	0.055	0.645	0.3	0
K-means on	100	10	0	0.785	0.09	0.085	0.04	0	0.835	0.065	0.085	0.015
eigenvectors	100	20	0	0.795	0.075	0.09	0.04	0	0.77	0.055	0.155	0.02
	100	40	0	0.82	0.085	0.08	0.015	0	0.78	0.075	0.105	0.04
	200	10	0	0.92	0.03	0.035	0.015	0	0.93	0.025	0.045	0
	200	20	0	0.87	0.05	0.06	0.02	0	0.87	0.055	0.065	0.01
	200	40	0	0.805	0.075	0.085	0.035	0	0.86	0.075	0.055	0.01
SBSA 1	100	10	0	0.11	0.65	0.225	0.015	0	0.06	0.795	0.135	0.01
	100	20	0	0	0.955	0.045	0	0	0	0.995	0.005	0
	100	40	0	0	1	0	0	0	0	1	0	0
	200	10	0	0	0.755	0.21	0.035	0	0.005	0.98	0.015	0
	200	20	0	0	0.985	0.015	0	0	0	1	0	0
	200	40	0	0	1	0	0	0	0	1	0	0
SBSA 2	100	10	0	0.005	0.995	0	0	0	0	0.995	0.005	0
	100	20	0	0	1	0	0	0	0	1	0	0
	100	40	0	0	1	0	0	0	0	1	0	0
	200	10	0	0	1	0	0	0	0	1	0	0
	200	20	0	0	1	0	0	0	0	1	0	0
	200	40	0	0	1	0	0	0	0	1	0	0

as T increases.

6 Empirical application

6.1 The model and data

Individual portfolio choices are influenced by many factors, some of which are observable and others are unobservable. For example, age, financial assets, labor income, and returns and risk measures of different assets are among the set of observable factors. For a seminal paper on the problem of portfolio choice, see Samuelson (1969). Cocco, Gomes, and Maenhout (2005) investigate how labor income and financial wealth affect portfolio decisions. Unobservable factors also play a very important role in the process of portfolio decision making. For example, individual risk preference, habits and information acquirement affect how people respond to various observable factors. Samuelson (1969) models risk preference as the fundamental factor in portfolio choices. Polkovnichenko (2007) employs n the life cycle model to study the implications of endogenous habit formation preferences on portfolio choices. Both academic studies and common sense suggest that different people tend to have different responses to the same information. This fact motivates us to consider the panel structure model in studying how individuals' portfolio choices are affected by various factors.

Table 3: Classification and point estimation of $\boldsymbol{\alpha}_{\cdot 2}^0$

			DGP 1				DGP 2			
			Correct	Comparison Criteria			Correct	Comparison Criteria		
	N	T	Ratio	RMSE	Bias	Coverage	Ratio	RMSE	Bias	Coverage
Oracle	100	10	1	0.059	0.000	0.933				
	100	20	1	0.041	-0.001	0.920	1	0.041	0.001	0.939
	100	40	1	0.028	-0.000	0.947	1	0.028	0.002	0.954
	200	10	1	0.040	-0.000	0.931				
	200	20	1	0.027	-0.002	0.950	1	0.028	-0.000	0.949
	200	40	1	0.019	0.001	0.947	1	0.020	0.001	0.948
K-means	100	10	0.917	0.289	-0.080	0.766				
on $\widetilde{\beta}$	100	20	0.964	0.276	-0.097	0.795	0.991	0.181	-0.029	0.897
	100	40	0.984	0.214	-0.057	0.873	0.988	0.197	-0.025	0.903
	200	10	0.920	0.260	-0.080	0.763				
	200	20	0.962	0.281	-0.097	0.831	0.991	0.134	-0.025	0.921
	200	40	0.975	0.248	-0.081	0.845	0.987	0.247	-0.033	0.892
K-means	100	10	0.794	0.524	-0.241	0.332				
on eigen-	100	20	0.861	0.381	-0.179	0.487	0.987	0.119	-0.005	0.891
vectors	100	40	0.955	0.278	-0.107	0.774	0.997	0.172	0.008	0.938
	200	10	0.784	0.540	-0.237	0.308				
	200	20	0.858	0.384	-0.180	0.425	0.989	0.070	-0.009	0.910
	200	40	0.944	0.332	-0.157	0.723	0.998	0.085	-0.002	0.942
C-Lasso	100	10	0.939	0.076	-0.017	0.866				
	100	20	0.985	0.044	-0.005	0.905	1	0.041	0.001	0.939
	100	40	0.999	0.028	-0.001	0.944	1	0.028	0.002	0.954
	200	10	0.941	0.052	-0.018	0.840				
	200	20	0.986	0.028	-0.005	0.942	1	0.028	-0.000	0.949
	200	40	0.999	0.019	0.000	0.943	1	0.020	0.001	0.948
SBSA 1	100	10	0.929	0.104	0.003	0.846				
	100	20	0.983	0.044	-0.002	0.903	0.791	0.504	-0.042	0.334
	100	40	0.999	0.028	-0.000	0.946	0.855	0.268	-0.023	0.327
	200	10	0.933	0.051	0.004	0.860				
	200	20	0.985	0.028	-0.001	0.941	0.778	0.482	-0.044	0.314
	200	40	0.999	0.019	0.001	0.946	0.852	0.226	-0.025	0.295
SBSA 2	100	10	0.931	0.076	0.004	0.856				
	100	20	0.984	0.043	-0.001	0.908	0.991	0.045	0.002	0.913
	100	40	0.999	0.028	-0.001	0.946	1	0.028	0.002	0.954
	200	10	0.931	0.050	0.006	0.864				
	200	20	0.984	0.029	-0.001	0.933	0.992	0.032	0.000	0.921
	200	40	0.999	0.019	0.001	0.946	1	0.020	0.001	0.948

Table 4: Classification and point estimation of $\boldsymbol{\alpha}_{\cdot 2}^0$

			DGP 3				DGP 4			
			Correct	Comparison Criteria			Correct	Comparison Criteria		
	N	T	Ratio	RMSE	Bias	Coverage	Ratio	RMSE	Bias	Coverage
Oracle	100	10	1	0.086	0.003	0.957	1	0.072	0.001	0.939
	100	20	1	0.063	0.003	0.947	1	0.048	0.002	0.958
	100	40	1	0.044	0.005	0.936	1	0.036	0.001	0.940
	200	10	1	0.067	0.003	0.928	1	0.058	-0.004	0.912
	200	20	1	0.045	0.002	0.931	1	0.040	-0.002	0.919
	200	40	1	0.031	0.001	0.946	1	0.030	-0.004	0.942
K-means	100	10	0.930	0.210	-0.020	0.839	0.900	0.261	-0.069	0.637
on $\widetilde{\beta}$	100	20	0.970	0.138	-0.010	0.833	0.933	0.273	-0.085	0.673
	100	40	0.985	0.163	-0.016	0.878	0.942	0.291	-0.112	0.623
	200	10	0.937	0.153	0.001	0.823	0.900	0.235	-0.064	0.626
	200	20	0.971	0.146	-0.011	0.825	0.935	0.222	-0.061	0.650
	200	40	0.982	0.150	-0.021	0.861	0.940	0.204	-0.054	0.696
K-means	100	10	0.751	0.316	-0.089	0.157	0.768	0.312	-0.067	0.199
on eigen-	100	20	0.764	0.355	-0.137	0.143	0.773	0.273	-0.068	0.211
vectors	100	40	0.768	0.368	-0.140	0.105	0.777	0.226	-0.064	0.168
	200	10	0.748	0.331	-0.107	0.100	0.757	0.293	-0.063	0.184
	200	20	0.758	0.287	-0.094	0.082	0.765	0.217	-0.064	0.181
	200	40	0.753	0.326	-0.118	0.033	0.767	0.230	-0.057	0.197
SBSA 1	100	10	0.885	0.180	0.062	0.541	0.877	0.124	0.017	0.650
	100	20	0.959	0.087	0.024	0.832	0.949	0.069	0.006	0.845
	100	40	0.991	0.048	0.009	0.920	0.982	0.040	0.002	0.914
	200	10	0.885	0.166	0.084	0.472	0.878	0.117	0.020	0.598
	200	20	0.962	0.074	0.031	0.755	0.956	0.058	0.003	0.770
	200	40	0.992	0.034	0.005	0.930	0.985	0.033	-0.001	0.926
SBSA 2	100	10	0.935	0.114	0.039	0.872	0.925	0.085	0.012	0.893
	100	20	0.986	0.065	0.006	0.940	0.972	0.051	0.004	0.944
	100	40	0.997	0.045	0.006	0.938	0.994	0.036	0.002	0.940
	200	10	0.936	0.097	0.044	0.831	0.929	0.063	0.010	0.881
	200	20	0.986	0.048	0.009	0.920	0.977	0.041	0.001	0.901
	200	40	0.998	0.031	0.002	0.948	0.997	0.030	-0.002	0.941

In this application, we consider a censored model similar to that in Abrevaya and Shen (2014, hereafter AS). The dependent variable y_{it} is the ratio of safe assets in individual i's portfolio in year t, and it is left censored at 0 and right censored at 1. To account for parameter heterogeneity, we consider the mixed panel structure model of the form

$$y_{it}^* = x_{1,it}^\top \beta_{1i} + x_{2,it}^\top \beta_2 + \mu_i + \varepsilon_{it}, \tag{6.1}$$

where $x_{1,it}$ includes log financial assets and log non-capital income, $x_{2,it}$ includes AEX premium, time trend and retirement dummy, μ_i is the fixed effect, and ε_{it}'s are i.i.d. normal.[4] The observable dependent variable y_{it} is subject to two-sided censoring: $y_{it} = \text{mami}\{0, y_{it}^*, 1\}$. Note that β_2 is common across individuals in (6.1). We assume that the true values of β_{1i}'s exhibit the group structure, $\beta_{1i}^0 = \sum_{k=1}^{K^0} \alpha_k^0 \cdot \mathbf{1}\left\{i \in G_k^0\right\}$. We are interested in identifying the number of groups (K^0) and the group membership for each individual i.

Next we explain briefly why we allow β_{1i}'s to be heterogeneous across groups and impose homogeneity assumption on β_2. The variables contained in $x_{1,it}$, namely, log financial asset and log non-capital income, are usually modeled as determinant factors in portfolio choice theories. Curcuru et al. (2004) argue that there is substantial heterogeneity in the portfolio choices. In other words, different people tend to have different responses towards the same factors. But individuals' behavior also tends to exhibit certain grouped patterns. For example, some individuals prefer to holding diversified portfolios in order to hedge against various kinds of risks whereas others hold almost no position on risky or riskless assets. In modeling economic behavior, homogeneous representative individual assumption is a convenient way to explain some phenomenon. But it is quite fragile as heterogeneity is ubiquitous. The panel structure model studied in this paper offers a flexible and manageable alternative to handle the parameter heterogeneity issue.

The retirement dummy, which is contained in $x_{2,it}$, may change over the time span for some individuals, and remains as a constant (0 or 1) for other individuals. To avoid the multicollinearity issue, we treat its coefficient as constant across i. Classic theory (e.g., Cocco et al. (2005)) generally predicts that the ratio of savings in safe assets tends to increase after retirement. AEX premium is believed to be negatively correlated to y_{it}, the ratio of safe asset in the portfolio. There are few reasons to believe otherwise. Besides, AS's regression results are aligned with these theoretical predictions, which motivates us to assume homogeneous effects of the variables in $x_{2,it}$ across individuals.

The dataset comes from the De Nederlandsche Bank (DNB) Household Survey of Netherlands, which contains detailed demographic and financial information of Dutch household and individual samples from 1993 to 2015. We use unbalanced panel data and first include all individuals with time dimension T_i larger than or equal to 10. There are $N = 378$ individuals included in our regression. The average period of observations for all individuals is about $N^{-1}\sum_{i=1}^N T_i \approx 12.3$.

[4] AEX premium is defined as Amsterdam exchange index return minus the deposit rate. The retirement age in Netherlands is 65. For detailed explanation of all variables defined here, please refer to Alessie, Hochguertel, and Van Soest (2002) and AS.

The majority of censoring is right censoring at one. To be specific, the right censored ratio is 1691 out of 4666 (36.2%); and the left censored ratio is 142 out of 4666 (3.0%). Table 5 provides a brief summary of the dataset.

Table 5: Summary statistics for the DNB household survey dataset

	y_{it}	log(FA)	log(NCI)	AEX prem.	Time (t)	Retire dummy
min.	0.0000	1.609	5.247	-0.475	2.000	0.000
max.	1.0000	14.881	13.768	0.384	23.000	1.000
mean	0.6606	9.852	10.227	0.009	13.012	0.260
median	0.8126	9.974	10.296	0.080	13.000	0.000
std.	0.3656	1.695	0.749	0.217	6.050	0.439

6.2 Classification and post-classification regression results

We apply our SBSA method to the above dataset and obtain the classification and post-classification regression results. Based on SBSA 2, IC_2 in (5.1) determines three estimated groups with Groups 1–3 containing 112, 100, and 166 individuals, respectively.

Table 6 reports the regression results for different specifications. Column (1) corresponds to the usual pooled censored panel data regression with fixed effects. Columns (2)–(4) correspond to the joint estimation of group-specific parameters and the common parameters in the model. Note that we assume the effects of variables in $x_{2,it}$ and the variance of the error terms are common across all individuals for this joint estimation. Column (5) collects some regression results, corresponding to the relevant variables used here, from AS. Following AS, we include many common explanatory variables and also use the censored regression model. That being said, the data used here are different from theirs. They use the DNB household survey from 1993 to 2008 with individuals' time periods (T_i) larger than or equal to three. Our data come from the same source, but range from 1993 to 2015 with individuals' time periods longer than or equal to ten.

We summarize some important findings from Table 6. First, the coefficient of log financial assets (log(FA)) is very similar between the pooled model (column (1)) and AS's model (column (5)). The negative relationship between log(FA) and safe asset ratio (y_{it}) is very stable across time and individuals. For the other regressors, our pooled estimates are somewhat different from those of AS's. The coefficient of the time trend is positive and significant at the 1% level while it is negative and significant at the 1% level in AS. One possible explanation is that we use data from individuals with periods of observation more than or equal to ten, which is longer than that of AS's. After many periods of portfolio decisions, a person gets older and older and tends to allocate more assets to safe investments. If the time periods are very short (three in AS's data for many individuals), the effect may not be captured properly. In short, when we choose to include individuals with periods of observations greater than or equal to 10, we tend to choose different samples than that of AS. It has some impacts on our regression results.

Second, our SBSA 2 method yields three estimated groups whose regression outputs are re-

Table 6: Regression results for the DNB household survey dataset

	(1) Pooled	(2) Group 1	(3) Group 2	(4) Group 3	(5) AS
log(FA)	-0.128^c	-0.055^c	-0.223^c	-0.048^c	-0.129^c
	(0.005)	(0.009)	(0.009)	(0.008)	(0.011)
log(NCI)	0.035^c	-0.255^c	0.056^c	0.091^c	-0.006^a
	(0.012)	(0.024)	(0.018)	(0.016)	(0.004)
AEX premium	0.008		-0.007		-0.039^b
	(0.023)		(0.022)		(0.017)
Time (t)	0.024^c		0.020^c		-0.013^c
	(0.001)		(0.001)		(0.002)
Retirement dummy	0.079^c		0.065^c		
	(0.021)		(0.020)		
σ^2	0.310^c		0.290^c		
	(0.004)		(0.004)		

Note: Column (1) reports the pooled estimation of all 378 individuals. By using SBSA 2, we obtain 3 groups. Columns (2)–(4) report the regression results for each group where the coefficients of AEX premium, time trend and retirement dummy are common. Column (5) reports part of regression results drawn from AS for comparison purpose. Standard errors are in parentheses. ^{a}, ^{b}, and ^{c} denote significance at 10%, 5% and 1% levels, respectively.

ported in Columns (2), (3), and (4) in Table 6. The table suggests that the signs of the coefficient estimates for log non-capital income (log(NCI)) are opposite for Group 1 and the other two groups while the signs of the coefficient estimates for log(FA) are common across all three groups. The former finding provides partial explanation for the opposite direction of log(NCI) in columns (1) and (6). There are three latent groups. Pooling them together yields a weighted average of the estimates in columns (2)–(4), which is positive for log(NCI) in column (1). Different composition of elements from the three groups might generate a negative slope for log(NCI) in the pooled estimation, e.g., in AS (column (6)).

Third, the effects of log(FA) on the ratio of safe assets (y_{it}) are similar in Groups 1 and 3 and they are much smaller than that in Group 2. So the separation between Groups 1 and 3 is mainly caused by the quite distinct effects of log(NCI) on the ratio of safe assets.

Fourth, our estimate of the common coefficient of AEX premium is negative, which is different from the pooled estimate but consistent with AS's results and the theoretical prediction.

6.3 Robustness check

In the above subsection we consider the classification and post-classification regression results by using SBSA 2 for individuals with $T_i \geq 10$. There are 378 individuals in total. As a robustness check, we now consider the cases where $T_i \geq 9$ or $T_i \geq 8$.

First, we consider the classification results based on individuals with $T_i \geq 9$. Now the number of individuals (N) increases to 504. By using the SBSA 2 method, we still obtain 3 groups.

Groups 1–3 contain 129, 121, and 254 individuals, respectively. The left panel of Table 7 reports the post-classification regression results in this case. A comparison with Table 6 suggests that the post-classification results share some similar patterns, in terms of both estimated number of groups and coefficient estimates for each group.

Next, we consider individuals with $T_i \geq 8$. There are 627 individuals for this case. We apply SBSA 2 method on this new subsample. As before, we obtain 3 groups. Groups 1–3 contain 116, 182, and 329 individuals, respectively. The post-classification regression results are reported in the right panel of Table 7. A comparison between Table 6 and the right panel of Table 7 suggests that the post-classification results here are similar to those in Table 6

In sum, we conclude that our SBSA 2 classification and estimation results are quite robust to the choice of the minimum value of T_i.

Table 7: Regression results for the DNB household survey data for $T_i \geq 9$ or 8 after using SBSA 2

	$T_i \geq 9$			$T_i \geq 8$		
	Group 1	Group 2	Group 3	Group 1	Group 2	Group 3
log(FA)	-0.043^c	-0.240^c	-0.055^c	-0.040^c	-0.224^c	-0.028^c
	(0.008)	(0.009)	(0.007)	(0.008)	(0.007)	(0.005)
log(NCI)	-0.304^c	0.027^c	0.068^c	-0.414^c	0.031^c	0.028^c
	(0.024)	(0.017)	(0.013)	(0.025)	(0.013)	(0.011)
AEX premium		-0.013			-0.010	
		(0.020)			(0.017)	
Time (t)		0.019^c			0.022^c	
		(0.001)			(0.001)	
Retirement dummy		0.069^c			0.052^c	
		(0.018)			(0.015)	
σ^2		0.290^c			0.266^c	
		(0.004)			(0.003)	

We might also want to know how many individuals in Group 1 when $T_i \geq 10$ are still in Group 1 when $T_i \geq 9$. Such statistics are reported in Table 8. For example, the number 0.857 in row 2 and column 2 in the table means that 85.7% of the members in Group 1 are still in Group 1 when we relax $T_i \geq 10$ to $T_i \geq 9$. Similarly, Table 9 reports the group membership shifts when the minimum T_i decreases from 9 to 8. Both Tables 8 and 9 show that the majority of individuals have stable membership when we decrease the minimum T_i.

Table 8: The classification membership shifts when minimum T_i changes from 10 to 9

Ratio	Group 1, $T_i \geq 10$	Group 2, $T_i \geq 10$	Group 3, $T_i \geq 10$
Group 1, $T_i \geq 9$	0.857	0	0
Group 2, $T_i \geq 9$	0.045	0.870	0
Group 3, $T_i \geq 9$	0.098	0.130	1.000

Table 9: The classification membership shifts when minimum T_i changes from 9 to 8

Ratio	Group 1, $T_i \geq 9$	Group 2, $T_i \geq 9$	Group 3, $T_i \geq 9$
Group 1, $T_i \geq 8$	0.674	0	0
Group 2, $T_i \geq 8$	0.109	1.000	0.067
Group 3, $T_i \geq 8$	0.217	0	0.933

7 Conclusion

In this paper we propose a sequential binary segmentation algorithm (SBSA) to estimate a panel structure model. This method is motivated from the intuition that the parameter heterogeneity problem can be translated into the break detection problem, which is well studied and understood in the time series literature. We also propose an information criterion to determine the number of groups. We show that our method can recover the true group structure w.p.a.1 and our post-classification estimators are oracally efficient. Furthermore, we build the link between the panel structure model and the stochastic block model (SBM) in the network literature. The linkage enables us to use community detection techniques from the SBM to the panel structure model. We apply SBSA on the eigenvectors corresponding to the few largest eigenvalues of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^{\top}$ and improve the finite sample performance significantly in some cases. Our method is easy to implement and efficient to compute. Simulations demonstrate superb finite sample performance of our method. We also apply our method to study how financial assets and non capital income, among others, affect individuals' portfolio choices by allowing unobserved parameter heterogeneity and using the DNB household survey dataset. We detect three latent groups in the dataset.

There are several possible extensions. First, we can also include time effects in our model. Following the asymptotic analysis of Chen (2016) we can also show that the preliminary estimates of the individual parameters are still $\sqrt{T}$-consistent, which enables us to conduct the SBSA as in the current paper to detect possible grouped patterns. Second, we do not allow cross sectional dependence in this paper. Chen et al. (2014) study homogeneous nonlinear panel data models with interactive fixed effects (IFEs) and Su and Ju (2017) consider a linear panel structure model with IFEs. It is possible to combine the approaches in these papers and study heterogeneous nonlinear panel data models with IFEs. Again, as long as we can establish the $\sqrt{T}$-consistency of the preliminary estimates of the individual parameters of interest, we can apply the SBSA to detect latent groups among them. Third, we do not allow nonstationary ($I(1)$) regressors in our model. It is possible to extend our method to nonstationary panels with latent group structures. Fourth, it is also possible to allow for structural changes in the model; see, e.g., Okui and Wang (2017). We leave these topics for future research.

APPENDIX

In this appendix we first state and prove some technical lemmas, and then prove the main results in the paper.

A Some technical lemmas

In this section we state and prove several technical lemmas that are used in the proofs of the main results in the paper.

Lemma A.1. *Let $\xi(w_{it};\varsigma)$ be a $\mathbb{R}^{d_\xi}$-valued function indexed by the parameter $\varsigma \in \Upsilon$, where Υ is a convex compact set in $\mathbb{R}^{d_\varsigma}$ and $\mathbb{E}[\xi(w_{it};\varsigma)] = 0$ for all i, t, and $\varsigma \in \Upsilon$. Assume that there exists a function $M(w_{it})$ such that $\|\xi(w_{it};\varsigma) - \xi(w_{it};\varsigma')\| \leq M(w_{it})\|\varsigma - \varsigma'\|$ for all ς, $\varsigma' \in \Upsilon$ and $\sup_\varsigma \|\xi(w_{it};\varsigma)\| \leq M(w_{it})$. Assume that $\mathbb{E}[M(w_{it})]^\kappa < \infty$ for some $\kappa \geq 6$ such that $N = O(T^{\kappa/2-1})$. Let $\{\varsigma_i\}$ be a nonstochastic sequence in Υ. Then $\max_i \|T^{-1/2}\sum_{t=1}^T \xi(w_{it};\varsigma_i)\| = O_P((\ln T)^3)$.*

Proof: This is Lemma S1.2(i) in SSPb. ■

Recall that $\gamma_i = (\beta_i^\top, \mu_i^\top)^\top$, $\gamma_i^0 = (\beta_i^{0\top}, \mu_i^{0\top})^\top$, and $\tilde{\gamma}_i = (\tilde{\beta}_i^\top, \tilde{\mu}_i^\top)^\top$. Let $\hat{\Psi}_i(\gamma,\theta) = \frac{1}{T}\sum_{t=1}^T \varphi(w_{it};\gamma,\theta)$ and $\Psi_i(\gamma,\theta) = \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\varphi(w_{it};\gamma,\theta)]$. Then $\gamma_i(\theta) = (\beta_i(\theta)^\top, \mu_i(\theta)^\top)^\top = \arg\min_{\gamma_i} \Psi_i(\gamma_i,\theta)$ and $\tilde{\gamma}_i(\theta) \equiv (\tilde{\beta}_i(\theta)^\top, \tilde{\mu}_i(\theta)^\top)^\top \equiv \arg\min_{\gamma_i} \hat{\Psi}_i(\gamma_i,\theta)$.

Lemma A.2. *Suppose that Assumption A1 holds. Then for any fixed $\eta > 0$ and $v > 0$, we have*
(i) $P\left(\max_i \sup_{(\gamma,\theta)} \left|\hat{\Psi}_i(\gamma,\theta) - \Psi_i(\gamma,\theta)\right| \geq \eta\right) = o(N^{-1})$,
(ii) $\tilde{\gamma}_i(\theta) - \gamma_i(\theta) = O_P(T^{-1/2})$ *for each* i,
(iii) $P\left(\max_i \sup_\theta \|\tilde{\gamma}_i(\theta) - \gamma_i(\theta)\| \geq \eta T^{-1/2}(\ln T)^{3+v}\right) = o(N^{-1})$,
(iv) $P\left(\max_i \sup_\theta \left|\frac{1}{N}\sum_{i=1}^N [\Psi_i(\tilde{\gamma}_i(\theta),\theta) - \Psi_i(\gamma_i(\theta),\theta)]\right| \geq \eta T^{-1/2}(\ln T)^{3+v}\right) = o(N^{-1})$,
(v) $\frac{1}{N}\sum_{i=1}^N \left\|\tilde{\gamma}_i(\theta^0) - \gamma_i(\theta^0)\right\|^2 = O_P(T^{-1})$.

Proof: (i), (ii) and (iii) follow from Lemmas S1.3, S1.5(i) and S1.5(iv) in SSPb by the repeated use of Lemma A.1 with little modifications. Noting that

$$\left|\frac{1}{N}\sum_{i=1}^N [\Psi_i(\tilde{\gamma}_i(\theta),\theta) - \Psi_i(\gamma_i(\theta),\theta)]\right| \leq \max_{i,t} \mathbb{E}[M(w_{it})] \frac{1}{N}\sum_{i=1}^N \|\tilde{\gamma}_i(\theta) - \gamma_i(\theta)\|$$

and $\max_{i,t} \mathbb{E}[M_i(w_{it})] \leq c_M^{1/\kappa}$ by Assumption A1(iv) and the Jensen inequality, (iv) follows from (iii). We are left to show (v). Recall that $Z(w_{it};\gamma_i,\theta) = \partial\varphi(w_{it};\gamma_i,\theta)/\partial\gamma_i$ and $Z^{\gamma_i}(w_{it};\gamma_i,\theta) = \partial Z(w_{it};\gamma_i,\theta)/\partial\gamma_i^\top$. Noting that $\tilde{\gamma}_i(\theta) = \arg\min_{\gamma_i} \hat{\Psi}_i(\gamma_i,\theta)$, we have

$$\begin{aligned} 0 &= \frac{1}{T}\sum_{t=1}^T Z(w_{it};\tilde{\gamma}_i(\theta),\theta) \\ &= \frac{1}{T}\sum_{t=1}^T Z(w_{it};\gamma_i(\theta),\theta) + \frac{1}{T}\sum_{t=1}^T \hat{H}_i(\theta) Z^{\gamma_i}(w_{it};\breve{\gamma}_i(\theta),\theta)[\tilde{\gamma}_i(\theta) - \gamma_i(\theta)], \end{aligned}$$

where $\hat{H}_i(\theta) = \frac{1}{T}\sum_{t=1}^T Z^{\gamma_i}(w_{it};\breve{\gamma}_i(\theta),\theta)$ and $\breve{\gamma}_i(\theta)$ lies between $\tilde{\gamma}_i(\theta)$ and $\gamma_i(\theta)$ elementwise. Then

$$\tilde{\gamma}_i(\theta^0) - \gamma_i(\theta^0) = -\hat{H}_i(\theta^0)^{-1}\frac{1}{T}\sum_{t=1}^T Z\left(w_{it};\gamma_i(\theta^0),\theta^0\right)$$

provided that $\hat{H}_i(\theta^0)$ is asymptotically nonsingular. Let $\hat{H}_i = \hat{H}_i(\theta^0)$, $\breve{H}_i = \frac{1}{T}\sum_{t=1}^{T} Z^{\gamma_i}(w_{it};\gamma_i(\theta^0),\theta^0)$, and $H_i = \mathbb{E}(\breve{H}_i)$. Under Assumption A1, we can readily show that

$$\max_{1\le i\le N}\|\hat{H}_i - H_i\| \le \max_{1\le i\le N}\|\hat{H}_i - \breve{H}_i\| + \max_{1\le i\le N}\|\breve{H}_i - H_i\| = o_P(1),$$

which implies that $\lambda_{\min}(\hat{H}_i) = \lambda_{\min}(H_i) + o_P(1)$ uniformly in i. Consequently, we have

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\gamma}_i(\theta^0) - \gamma_i(\theta^0)\right\|^2 &\le \left[\min_{1\le i\le N}\lambda_{\min}(H_i) + o_P(1)\right]^{-1}\frac{1}{NT^2}\sum_{i=1}^{N}\left\|\sum_{t=1}^{T} Z\left(w_{it};\gamma_i(\theta^0),\theta^0\right)\right\|^2 \\ &= O_P(1)\,O_P(T^{-1}) = O_P(T^{-1}),\end{aligned}$$

where we use the fact that $\mathbb{E}[Z\left(w_{it};\gamma_i(\theta^0),\theta^0\right)] = 0$ and that $\mathbb{E}\|\sum_{t=1}^{T} Z(w_{it};\gamma_i(\theta^0),\theta^0)\|^2 = O(T)$ by a simple application of the Davydov inequality under Assumption A1 (e.g., Hall and Heyde (1980, p. 278)). ■

Lemma A.3. *Suppose that Assumption A1 holds. Then* $P(\|\tilde{\theta} - \theta^0\| > \eta) = o(N^{-1})$.

Proof: Noting that $\tilde{\theta} = \arg\min_\theta Q_{NT}(\theta)$, we have $Q_{NT}(\tilde{\theta}) \le Q_{NT}(\theta^0)$. By Assumption A1(iv), there exists a constant $\epsilon > 0$ such that $\inf_{\theta:\|\theta-\theta^0\|>\eta}\frac{1}{N}\sum_{i=1}^{N}\left[\Psi_i\left(\gamma_i\left(\theta\right),\theta\right) - \Psi_i(\gamma_i(\theta^0),\theta^0)\right] \ge \epsilon$. Define

$$\begin{aligned}A_1 &\equiv \left\{\max_{1\le i\le N}\sup_{(\gamma_i,\theta)}\left|\hat{\Psi}_i\left(\gamma_i,\theta\right) - \Psi_i\left(\gamma_i,\theta\right)\right| \le \frac{1}{6}\epsilon\right\}, \text{ and} \\ A_2 &\equiv \left\{\sup_{\theta\in\Theta}\left|\frac{1}{N}\sum_{i=1}^{N}\left[\Psi_i\left(\tilde{\gamma}_i\left(\theta\right),\theta\right) - \Psi_i\left(\gamma_i\left(\theta\right),\theta\right)\right]\right| \le \frac{1}{6}\epsilon\right\}.\end{aligned}$$

By Lemma A.2(i) and (iii)–(iv) and Assumption A1(ii), $P\left(A_1 \cap A_2\right) \ge 1 - P\left(A_1^c\right) - P\left(A_2^c\right) = 1 - o(N^{-1})$. Then conditional on $A_1 \cap A_2$, we have

$$\begin{aligned}\inf_{\theta:\|\theta-\theta^0\|>\eta}\frac{1}{N}\sum_{i=1}^{N}\hat{\Psi}_i\left(\tilde{\gamma}_i\left(\theta\right),\theta\right) &\ge \inf_{\theta:\|\theta-\theta^0\|>\eta}\frac{1}{N}\sum_{i=1}^{N}\Psi_i\left(\tilde{\gamma}_i\left(\theta\right),\theta\right) - \frac{1}{6}\epsilon \\ &\ge \inf_{\theta:\|\theta-\theta^0\|>\eta}\frac{1}{N}\sum_{i=1}^{N}\Psi_i\left(\gamma_i\left(\theta\right),\theta\right) - \frac{1}{6}\varepsilon - \frac{1}{6}\epsilon \\ &\ge \frac{1}{N}\sum_{i=1}^{N}\Psi_i\left(\gamma_i(\theta^0),\theta^0\right) + \epsilon - \frac{1}{6}\epsilon - \frac{1}{6}\epsilon \\ &\ge \frac{1}{N}\sum_{i=1}^{N}\Psi_i\left(\tilde{\gamma}_i(\theta^0),\theta^0\right) - \frac{1}{6}\epsilon + \epsilon - \frac{1}{6}\epsilon - \frac{1}{6}\epsilon \\ &\ge \frac{1}{N}\sum_{i=1}^{N}\hat{\Psi}_i\left(\tilde{\gamma}_i(\theta^0),\theta^0\right) - \frac{1}{6}\epsilon - \frac{1}{6}\epsilon + \epsilon - \frac{1}{6}\epsilon - \frac{1}{6}\epsilon \\ &= \frac{1}{N}\sum_{i=1}^{N}\hat{\Psi}_i\left(\tilde{\gamma}_i(\theta^0),\theta^0\right) + \frac{1}{3}\epsilon.\end{aligned}$$

On the other hand, $\frac{1}{N}\sum_{i=1}^{N}\hat{\Psi}_i(\tilde{\gamma}_i(\tilde{\theta}),\tilde{\theta}) \le \frac{1}{N}\sum_{i=1}^{N}\hat{\Psi}_i(\tilde{\gamma}_i(\theta^0),\theta^0)$. It follows that $P(\|\tilde{\theta} - \theta^0\| > \eta) = o(N^{-1})$. ■

To state and prove the next lemma, we follow Hahn and Newey (2004) and SSPb and introduce some notation. Let F_i and $\hat{F}_i$ denote the cumulative and empirical distribution functions of w_{it}, respectively. Let $F_i(\epsilon) \equiv F_i + \epsilon\sqrt{T}(\hat{F}_i - F_i)$ for $\epsilon \in \left[0, T^{-1/2}\right]$. For fixed θ and ϵ, let $\gamma_i(\theta, F_i(\epsilon)) \equiv$

$\arg\min_{\gamma_i} \int \psi(\cdot;\gamma_i,\theta)\, dF_i(\epsilon)$, which is the solution to the estimating equation

$$0 = \frac{1}{N}\sum_{i=1}^{N} \int Z_i(\cdot;\gamma_i(\theta, F_i(\epsilon)),\theta)\, dF_i(\epsilon).$$

Define $\gamma_i^\theta(\epsilon) = \partial\gamma_i(\theta, F_i(\epsilon))/\partial\theta^\top$. Apparently, $F_i(0) = F_i$, $F_i\left(T^{-1/2}\right) = \hat{F}_i$, $\gamma_i(\theta) = \gamma_i(\theta, F_i(0))$, $\tilde{\gamma}_i(\theta) = \gamma_i(\theta, F_i(T^{-1/2}))$, $\frac{\partial\gamma_i(\theta)}{\partial\theta^\top} = \frac{\partial\gamma_i(\theta,F_i(0))}{\partial\theta^\top} = \gamma_i^\theta(0)$, and $\frac{\partial\tilde{\gamma}_i(\theta)}{\partial\theta^\top} = \frac{\partial\gamma_i(\theta,F_i(T^{-1/2}))}{\partial\theta^\top} = \gamma_i^\theta(T^{-1/2})$. We study the properties of $\gamma_i(\theta, F_i(\epsilon))$ and $\gamma_i^\theta(\epsilon)$ in the next lemma.

Lemma A.4. *Suppose that Assumption A1 holds. Then*

(i) $P\left(\max_{1\le i\le N}\max_{0\le\epsilon\le T^{-1/2}} \|\gamma_i(\theta, F_i(\epsilon)) - \gamma_i(\theta)\| \ge \eta\right) = o(N^{-1})$ *for any* $\eta > 0$,

(ii) $\max_{1\le i\le N, \|\theta-\theta^0\|=o(1)} \|\gamma_i(\theta) - \gamma_i(\theta^0)\| = o(1)$,

(iii) $P\left(\max_{1\le i\le N, \|\theta-\theta^0\|=o(1)} \|\tilde{\gamma}_i(\theta) - \tilde{\gamma}_i(\theta^0)\| \ge \eta\right) = o(N^{-1})$ *for any* $\eta > 0$,

(iv) $P\left(\max_{1\le i\le N}\max_{0\le\epsilon\le T^{-1/2}} \|\frac{\partial\gamma_i(\theta,F_i(\epsilon))}{\partial\theta^\top} - \frac{\partial\gamma_i(\theta)}{\partial\theta^\top}\| \ge \eta\right) = o(N^{-1})$ *for any* $\eta > 0$,

(v) $\max_{1\le i\le N, \|\theta-\theta^0\|=o(1)} \|\frac{\partial\gamma_i(\theta)}{\partial\theta^\top} - \frac{\partial\gamma_i(\theta^0)}{\partial\theta^\top}\| = o(1)$,

(vii) $P\left(\max_{1\le i\le N, \|\theta-\theta^0\|=o(1)} \|\frac{\partial\tilde{\gamma}_i(\theta)}{\partial\theta^\top} - \frac{\partial\tilde{\gamma}_i(\theta^0)}{\partial\theta^\top}\| \ge \eta\right) = o(N^{-1})$ *for any* $\eta > 0$.

Proof: The proofs of (i)–(iii) parallel those of Lemma S1.8(i)–(iii) in SSPb and thus are omitted. Similarly, the proofs of (iv)–(vi) parallel those of Lemma S1.9(i)–(iii) in SSPb. ■

B Proof of the main results

Proof of Theorem 3.1: (i) Noting that $\tilde{\theta} = \arg\min_\theta Q_{NT}(\theta)$, by the second order Taylor expansion and the envelope theorem, we have

$$\begin{aligned} 0 \ge Q_{NT}(\tilde{\theta}) - Q_{NT}(\theta^0) &= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\Psi}_i(\tilde{\gamma}_i(\tilde{\theta}),\tilde{\theta}) - \hat{\Psi}_i(\tilde{\gamma}_i(\theta^0),\theta^0)\right] \\ &= \tilde{\delta}^\top\hat{S} + \frac{1}{2}\tilde{\delta}^\top\hat{H}(\breve{\theta})\tilde{\delta} \ge \frac{1}{2}\lambda_{\min}(\hat{H}(\breve{\theta}))\|\tilde{\delta}\|^2 - \|\hat{S}\|\cdot\|\tilde{\delta}\|, \end{aligned}$$

where $\tilde{\delta} = \tilde{\theta} - \theta^0$, $\breve{\theta}$ lies between $\tilde{\theta}$ and θ^0 elementwise, $\hat{S} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Z(w_{it};\tilde{\gamma}_i(\theta^0),\theta^0)$, and

$$\hat{H}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[Z^\theta(w_{it};\tilde{\gamma}_i(\theta),\theta) + Z^{\gamma_i}(w_{it};\tilde{\gamma}_i(\theta),\theta)\frac{\partial\tilde{\gamma}_i(\theta)}{\partial\theta^\top}\right].$$

It follows that $\|\tilde{\delta}\| \le 2[\lambda_{\min}(\hat{H}(\breve{\theta}))]^{-1}\|\hat{S}\|$. For $\hat{S}$, we have

$$\hat{S} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} Z(w_{it};\gamma_i(\theta^0),\theta^0) + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left[Z(w_{it};\tilde{\gamma}_i(\theta^0),\theta^0) - Z(w_{it};\gamma_i(\theta^0),\theta^0)\right] \equiv S_1 + S_2, \text{ say.}$$

Noting that $\mathbb{E}(S_1) = 0$ and $\mathrm{Var}(S_1) = O((NT)^{-1})$, we have $S_1 = O_P((NT)^{-1/2})$. For S_2, we have by the Cauchy-Schwarz and Markov inequalities, Assumption A1(iv), and Lemma A.2(v)

$$\begin{aligned} \|S_2\| &\le \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\|Z^{\gamma_i}(w_{it};\tilde{\gamma}_i(\theta),\theta)\|\cdot\|\tilde{\gamma}_i(\theta^0) - \gamma_i(\theta^0)\| \\ &\le \left[\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} M(w_{it})^2\right]^{1/2}\left[\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\gamma}_i(\theta^0) - \gamma_i(\theta^0)\right\|^2\right]^{1/2} \\ &= O_P(1)\, O_P(T^{-1/2}) = O_P(T^{-1/2}). \end{aligned}$$

Then $\hat{S} = O_P(T^{-1/2})$.

To study $\hat{H}(\breve{\theta})$, recall $\breve{H}(\theta) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[Z^{\theta}(w_{it};\gamma_i(\theta),\theta)+Z^{\gamma_i}(w_{it};\gamma_i(\theta),\theta)\frac{\partial\gamma_i(\theta)}{\partial\theta^{\top}}]$ and let $H(\theta) = \mathbb{E}[\breve{H}(\theta)]$. Then by the triangle inequality

$$\|\hat{H}(\breve{\theta}) - H(\theta^0)\| \leq \|\hat{H}(\breve{\theta}) - \hat{H}(\theta^0)\| + \|\hat{H}(\theta^0) - \breve{H}(\theta^0)\| + \|\breve{H}(\theta^0) - H(\theta^0)\|.$$

Following the proof of Lemma S1.10 in SSPb, we can readily apply Assumption A1 and the results in Lemmas A.3–A.4 to show that each term on the right hand of the last expression is $o_P(1)$. Consequently we have $\lambda_{\min}(\hat{H}(\breve{\theta})) = \lambda_{\min}(H(\theta^0)) - o_P(1)$ and $\|\tilde{\theta} - \theta^0\| = \|\tilde{\delta}\| \leq 2[\lambda_{\min}(H(\theta^0)) - o_P(1)]^{-1}\|\hat{S}\| = O_P(T^{-1/2})$.

(ii) Noting that $\tilde{\gamma}_i = \tilde{\gamma}_i(\tilde{\theta})$ where $\tilde{\gamma}_i(\theta) = \arg\min_{\gamma_i}\hat{\Psi}_i(\gamma_i,\theta)$, we have

$$\begin{aligned} 0 \geq \hat{\Psi}_i(\tilde{\gamma}_i,\tilde{\theta}) - \hat{\Psi}_i(\gamma_i^0,\tilde{\theta}) &= \frac{1}{T}\sum_{t=1}^{T}\left[\varphi(w_{it};\tilde{\gamma}_i(\tilde{\theta}),\tilde{\theta}) - \varphi(w_{it};\gamma_i^0,\tilde{\theta})\right] \\ &= \tilde{b}_i^{\top}\hat{S}_i + \frac{1}{2}\tilde{b}_i^{\top}H_{i,\theta\theta}(\breve{\gamma}_i,\breve{\theta})\tilde{b}_i + O_P(T^{-1/2}), \end{aligned}$$

where $\tilde{b}_i \equiv \tilde{\gamma}_i - \gamma_i^0$, $\breve{\gamma}_i$ lies between $\tilde{\gamma}_i$ and γ_i^0 elementwise, $\breve{\theta}$ lies between $\tilde{\theta}$ and θ^0 elementwise, $\tilde{\theta} - \theta^0 = O_P(T^{-1/2})$ from (i) above, $\hat{S}_i = \frac{1}{T}\sum_{t=1}^{T}Z(w_{it};\gamma_i^0,\theta^0)$, and

$$H_{i,\theta\theta}(\gamma_i,\theta) = \frac{1}{T}\sum_{t=1}^{T}\left[Z^{\theta}(w_{it};\gamma_i,\theta) + Z^{\gamma_i}(w_{it};\gamma_i,\theta)\frac{\partial\tilde{\gamma}_i(\theta)}{\partial\theta^{\top}}\right].$$

It follows that $\|\tilde{b}_i\| \leq 2[\lambda_{\min}(H_{i,\theta\theta}(\breve{\gamma}_i,\breve{\theta}))]^{-1}\|\hat{S}_i\| + O_P(T^{-1/2})$. As in the proof of Lemma A.2(v), we can readily show that $\hat{S}_i = O_P(T^{-1/2})$ and $\lambda_{\min}(H_{i,\theta\theta}(\breve{\gamma}_i,\breve{\theta})) = \lambda_{\min}(H_{i,\theta\theta}(\theta^0)) + o_P(1)$ uniformly in i. It follows that

$$\left\|\tilde{\gamma}_i - \gamma_i^0\right\| = \|\tilde{b}_i\| = O_P(T^{-1/2}).$$

(iii) By Lemma A.1, $\max_{1\leq i\leq N}\|\hat{S}_i\| = O_P(T^{-1/2}(\ln T)^3)$. This, in conjunction with the fact that $\lambda_{\min}(H_{i,\theta\theta}(\breve{\gamma}_i,\breve{\theta})) = \lambda_{\min}(H_{i,\theta\theta}(\theta^0)) + o_P(1)$ uniformly in i, implies that $\max_i\|\tilde{\gamma}_i - \gamma_i^0\| = O_P(T^{-1/2}(\ln T)^3)$.

(iv) $\frac{1}{N}\sum_{i=1}^{N}\left\|\tilde{\gamma}_i - \gamma_i^0\right\|^2 \leq 4[\min_{1\leq i\leq N}\lambda_{\min}(H_{i,\theta\theta}(\breve{\gamma}_i,\breve{\theta}))]^{-2}\frac{1}{N}\sum_{i=1}^{N}\|\hat{S}_i\|^2 = O_P(T^{-1})$ by the uniform consistency of $H_{i,\theta\theta}$ and the fact that $\frac{1}{N}\sum_{i=1}^{N}\|\hat{S}_i\|^2 = O_P(T^{-1})$. ■

Proof of Theorem 3.2: Let $u_i = \tilde{\beta}_i - \beta_i^0$. By Theorem 3.1, $u_i = O_P(T^{-1/2})$ and $\max_{1\leq i\leq N}\|u_i\| = O_P(T^{-1/2}(\ln T)^3)$. Without loss of generality (W.l.o.g.), we focus on the proof of the theorem when $K^0 = 3$ and then remark on the other cases. By ranking the preliminary estimates $\{\tilde{\beta}_i\}$ according to their jth elements, we have

$$\tilde{\beta}_{\pi_j(1),j} \leq \tilde{\beta}_{\pi_j(2),j} \leq \cdots \leq \tilde{\beta}_{\pi_j(N),j} \text{ for } j = 1,\ldots,p, \tag{B.1}$$

where $\{\pi_j(1),\ldots,\pi_j(N)\}$ is a permutation of $\{1,2,\ldots,N\}$ that is implicitly determined by the ranking relation in (B.1).

Let α_k^0 denote the true group-specific parameter value for Group k and $\alpha_{k,j}^0$ the jth element of α_k^0 for $j = 1,\ldots,p$ and $k = 1,\ldots,K^0$. For each regressor j, it falls into the three cases below:

Case 1: $\alpha_{1,j}^0 < \alpha_{2,j}^0 < \alpha_{3,j}^0$, $\alpha_{2,j}^0 < \alpha_{3,j}^0 < \alpha_{1,j}^0$, or $\alpha_{3,j}^0 < \alpha_{1,j}^0 < \alpha_{2,j}^0$ and so on. W.l.o.g., we will consider the subcase where $\alpha_{1,j}^0 < \alpha_{2,j}^0 < \alpha_{3,j}^0$ as the other subcases can be done through the relabeling of the group numbers.

Case 2: $\alpha_{1,j}^0 = \alpha_{2,j}^0 < \alpha_{3,j}^0$, $\alpha_{1,j}^0 < \alpha_{2,j}^0 = \alpha_{3,j}^0$, $\alpha_{2,j}^0 = \alpha_{3,j}^0 < \alpha_{1,j}^0$, $\alpha_{2,j}^0 < \alpha_{3,j}^0 = \alpha_{1,j}^0$, $\alpha_{3,j}^0 = \alpha_{1,j}^0 < \alpha_{2,j}^0$, or $\alpha_{3,j}^0 < \alpha_{1,j}^0 = \alpha_{2,j}^0$. W.l.o.g., we will analyze the subcase where $\alpha_{1,j}^0 = \alpha_{2,j}^0 < \alpha_{3,j}^0$ as similar analysis applies to the subcase where $\alpha_{1,j}^0 < \alpha_{2,j}^0 = \alpha_{3,j}^0$ and the other subcases through relabeling of the group numbers. Note that when $\alpha_{1,j}^0 = \alpha_{2,j}^0$, Groups 1 and 2 members are mixed in the ranking relation in (B.1) according to the jth regressor.

Case 3: $\alpha_{1,j}^0 = \alpha_{2,j}^0 = \alpha_{3,j}^0$. In this case, by using the ranking relation (B.1) we cannot separate any group from the others based on the jth regressor.

Let $\mathcal{S}_l$ denote the collection of the regressor indices such that the conditions in Case l are satisfied for $l = 1, 2, 3$. Apparently, $\mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3 = \{1, 2, \ldots, p\}$ and $\mathcal{S}_l \cap \mathcal{S}_{l'} = \emptyset$ for $l \neq l'$. Assumption A.2(i) ensures that $\mathcal{S}_1 \cup \mathcal{S}_2$ must be nonempty.

Let $\hat{\jmath}_1 = \arg\max_{1 \leq j \leq p} \tilde{V}_{1,N}(j)$ where $\tilde{V}_{1,N}(j) = \tilde{V}_{1,N}^0(j)/\bar{\sigma}_{1,N}^2(j)$. Apparently, $\bar{\sigma}_{1,N}^2(j)$ is bounded away from zero in probability for each j. By Theorem 3.1, the sample variance of $\{\tilde{\beta}_{i,j}, i = 1, \ldots, N\}$ converges to a positive constant c_j, say, for any $j \in \mathcal{S}_1 \cup \mathcal{S}_2$, whereas that of $\{\tilde{\beta}_{i,j'}, i = 1, \ldots, N\}$ is $O(T^{-1})$ for any $j' \in \mathcal{S}_3$. As a result, $P(\hat{\jmath}_1 \in \mathcal{S}_1 \cup \mathcal{S}_2) \to 1$ and index $\hat{\jmath}_1$ is chosen to estimate the break points in the whole sample $\mathcal{S}_{1,N}(j) \equiv \{\tilde{\beta}_{\pi_j(1),j}, \tilde{\beta}_{\pi_j(2),j}, \ldots, \tilde{\beta}_{\pi_j(N),j}\}$ in the first step of the SBSA for some $j \in \mathcal{S}_1 \cup \mathcal{S}_2$ such that $\hat{\jmath}_1 = j$. We will show no matter whether j is in $\mathcal{S}_1$ or $\mathcal{S}_2$, we can always identify one break point in $\mathcal{S}_{1,N}(j)$ w.p.a.1.

The second break point can be identified by choosing $j \in \mathcal{S}_1$ or $j \in \mathcal{S}_1 \cup \mathcal{S}_2$ depending on whether the break point $\{N_1 + N_2\}$ or $\{N_1\}$ is identified in the first step. For example, if in the first step we rely on some $j \in \mathcal{S}_1$ to identify the break point $\{N_1\}$ that distinguishes the first group from the rest two groups, then in the second step, we may rely on an element j from either $\mathcal{S}_1$ or $\mathcal{S}_2$ to identify the second break point $\{N_1 + N_2\}$ that separates the second group from the third group. On the other hand, if the break point $\{N_1 + N_2\}$ is identified in the first step to separate the third group from the rest two groups, then in the second step, we can only rely on some $j \in \mathcal{S}_1$ to identify the break point $\{N_1\}$ to separate the first and second groups. In this second case, we will show that $P(\hat{m}_1 = N_1 + N_2) \to 1$ as $(N, T) \to \infty$, which implies that w.p.a.1 $\hat{\jmath}_2 \equiv \arg\max_{1 \leq j \leq p}[\tilde{V}_{1,\hat{m}_1}(j) + \tilde{V}_{\hat{m}_1+1,N}(j)] = \arg\max_{1 \leq j \leq p}[\tilde{V}_{1,N_1+N_2}(j) + \tilde{V}_{N_1+N_2+1,N}(j)]$. Since the segment $\mathcal{S}_{N_1+N_2+1,N}(j)$ does not contain any break point, $\tilde{V}_{N_1+N_2+1,N}(j) = O_P\left(T^{-1}\right)$ for any $j \in \{1, 2, \ldots, p\}$ by Theorem 3.1. But $\tilde{V}_{1,N_1+N_2}(j)$ is bounded away from zero in probability for any $j \in \mathcal{S}_1$ and $O_P\left(T^{-1}\right)$ for any $j \in \mathcal{S}_2 \cup \mathcal{S}_3$. As a result, $P(\hat{\jmath}_2 \in \mathcal{S}_1) \to 1$ as $(N, T) \to \infty$. Our choice of selecting $\hat{\jmath}_{K-1}$ in the SBSA ensures that such an argument continues to hold when we have $K^0 > 3$ groups as long as the K^0 groups are separable from each other as required explicitly in Assumption A2(i).

Below, we prove that when either Case 1 or Case 2 applies (i.e., $\hat{\jmath}_1 \in \mathcal{S}_1$ or $\hat{\jmath}_1 \in \mathcal{S}_2$), we can successfully identify one break point in the first step of the SBSA. After one break point is identified in the first step, we can also identify the second break point in the second step no matter whether $\hat{\jmath}_2 \in \mathcal{S}_1$ or $\hat{\jmath}_2 \in \mathcal{S}_2$.

Case 1: $\hat{\jmath}_1 \in \mathcal{S}_1$. Based on the ranking relation in (B.1) and the fact that $\max_{1 \leq i \leq N} \|u_i\| = o_P(1)$, we have the following homogeneity property

$$\beta_{\pi_j(i)}^0 = \begin{cases} \alpha_1^0 & \text{if } 1 \leq i \leq N_1, \\ \alpha_2^0 & \text{if } N_1 + 1 \leq i \leq N_1 + N_2, \text{ for any } j \in \mathcal{S}_1. \\ \alpha_3^0 & \text{if } N_1 + N_2 + 1 \leq i \leq N, \end{cases}$$

Fix $j \in \mathcal{S}_1$ and $\hat{m}_1(j) = \arg\min_{1 \leq m < N} S_{1,N}(j, m)$. We consider three subcases:

Case 1a: $\frac{\tau_1}{\tau_1+\tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 > \frac{\tau_3}{\tau_2+\tau_3}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2$, ensuring $P(S_{1,N}(j, N_1 + N_2) - S_{1,N}(j, N_1) > 0) \to 1$.[5]

Case 1b: $\frac{\tau_1}{\tau_1+\tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 > \frac{\tau_3}{\tau_2+\tau_3}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2$, ensuring $P(S_{1,N}(j, N_1 + N_2) - S_{1,N}(j, N_1) < 0) \to 1$.

Case 1c: $\frac{\tau_1}{\tau_1+\tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 = \frac{\tau_3}{\tau_2+\tau_3}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2$.

W.l.o.g., we focus on Case 1a and will show that $P(\hat{m}_1(j) = N_1) \to 1$ as $(N, T) \to \infty$ by proving that (i) $P(\hat{m}_1(j) < N_1) \to 0$; (ii) $P(N_1 < \hat{m}_1(j) \leq N_1 + N_2) \to 0$; (iii) $P(N_1 + N_2 < \hat{m}_1(j) \leq N) \to 0$ as

[5] The condition can also be written as $\frac{\tau_1}{\tau_1+\tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 > \frac{1-\tau_1-\tau_2}{1-\tau_1}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2$, similar to equation (6) in Bai (1997).

$(N,T) \to \infty$. Then by mere symmetry, we can show that $P(\hat{m}_1(j) = N_1 + N_2) \to 1$ as $(N,T) \to \infty$ in Case 1b, and by arguments similar to those used in the proof of Lemma 7 in Bai (1997), we can show that $P(\hat{m}_1(j) = N_1) = P(\hat{m}_1(j) = N_1 + N_2) = \frac{1}{2}$. In each subcase, we can identify one break point $\{N_1\}$ or $\{N_1 + N_2\}$ in the first step of the SBSA.

We first show (i). When $m < N_1$, the average of $\tilde{\beta}_{i,j}$ over the two binary segments are

$$\bar{\beta}_{1,m}(j) = \alpha^0_{1,j} + \bar{u}_{1,m}(j) \text{ and } \bar{\beta}_{m+1,N}(j) = \frac{N_1 - m}{N - m}\alpha^0_{1,j} + \frac{N_2}{N-m}\alpha^0_{2,j} + \frac{N_3}{N-m}\alpha^0_{3,j} + \bar{u}_{m+1,N}(j),$$

where $\bar{u}_{1,m}(j) = m^{-1}\sum_{i=1}^{m} u_{\pi_j(i),j}$ and $\bar{u}_{m+1,N}(j) = (N-m)^{-1}\sum_{i=m+1}^{N} u_{\pi_j(i),j}$. Noting that $\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{1,m}(j) = u_{\pi_j(i),j} - \bar{u}_{1,m}(j)$ when $m < N_1$, we have

$$\sum_{i=1}^{m} |\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{1,m}(j)|^2 = \sum_{i=1}^{m} |u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2 = \sum_{i=1}^{m} |u_{\pi_j(i),j}|^2 - m|\bar{u}_{1,m}(j)|^2.$$

Similarly, noting that

$$\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{m+1,N}(j) = \begin{cases} a_{1m} + u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j) & \text{if } m+1 \le i \le N_1, \\ a_{2m} + u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j) & \text{if } N_1 + 1 \le i \le N_1 + N_2, \\ a_{3m} + u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j) & \text{if } N_1 + N_2 + 1 \le i \le N, \end{cases}$$

where $a_{1m} = (N-m)^{-1}[N_2(\alpha^0_{1,j} - \alpha^0_{2,j}) + N_3(\alpha^0_{1,j} - \alpha^0_{3,j})]$, $a_{2m} = (N-m)^{-1}[(N_1 - m)(\alpha^0_{2,j} - \alpha^0_{1,j}) + N_3(\alpha^0_{2,j} - \alpha^0_{3,j})]$, $a_{3m} = (N-m)^{-1}[(N_1 - m)(\alpha^0_{3,j} - \alpha^0_{1,j}) + N_2(\alpha^0_{3,j} - \alpha^0_{2,j})]$, and we suppress the dependence of a_{1m}, a_{2m} and a_{3m} on j, we have

$$\begin{aligned} \sum_{i=m+1}^{N} |\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{m+1,N}(j)|^2 &= (N_1 - m)|a_{1m}|^2 + N_2|a_{2m}|^2 + N_3|a_{3m}|^2 \\ &\quad + 2a_{1m}\sum_{i=m+1}^{N_1}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] + 2a_{2m}\sum_{i=N_1+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] \\ &\quad + 2a_{3m}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] + \sum_{i=m+1}^{N}|u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2. \end{aligned}$$

It follows that $S_{1,N}(j,m) = \frac{1}{N}\{\sum_{i=1}^{m}|\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{1,m}(j)|^2 + \sum_{i=m+1}^{N}|\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{m+1,N}(j)|^2\} = M_{1j}(m) + r_{1j}(m)$, where $M_{1j}(m) = \frac{N_1 - m}{N}|a_{1m}|^2 + \frac{N_2}{N}|a_{2m}|^2 + \frac{N_3}{N}|a_{3m}|^2$, and

$$\begin{aligned} r_{1j}(m) &= \frac{1}{N}\left[\sum_{i=1}^{m}|u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2 + \sum_{i=m+1}^{N}|u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2\right] \\ &\quad + \frac{2a_{1m}}{N}\sum_{i=m+1}^{N_1}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] + \frac{2a_{2m}}{N}\sum_{i=N_1+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] \\ &\quad + \frac{2a_{3m}}{N}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)]. \end{aligned}$$

Noting that a_{1m}, a_{2m}, and a_{3m} are $O(1)$ uniformly in $m \le N_1$, we can readily apply Theorem 3.1(iii) and show that $r_{1j}(m) = O_P(T^{-1/2}(\ln T)^3)$ uniformly in $m < N_1$. Now, observe that $S_{1,N}(j,m) - S_{1,N}(j,N_1) =$

$[M_{1j}(m) - M_{1j}(N_1)] + [r_{1j}(m) - r_{1j}(N_1)]$. By straightforward but tedious calculations, we can show that

$$\begin{aligned}\Delta M_{1j}(m) &\equiv M_{1j}(m) - M_{1j}(N_1) \\ &= \frac{N_1 - m}{N(1-m/N)(1-N_1/N)} \left| \left(1 - \frac{N_1}{N}\right)(\alpha^0_{1,j} - \alpha^0_{2,j}) + \left(1 - \frac{N_1+N_2}{N}\right)(\alpha^0_{2,j} - \alpha^0_{3,j}) \right|^2 \\ &\asymp \mu_{1mN} \text{ for all } m < N_1,\end{aligned}$$

where $\mu_{1mN} = (N_1 - m)/N$, and $a_N \asymp b_N$ denotes that both a_N/b_N and b_N/a_N converge to a positive number as $N \to \infty$. Note that $\lim_{N\to\infty} \mu_{1mN} \in [0, \tau_1]$ for each $m < N_1$. Specifically, $\mu_{1mN} \to \tau_1$ if $m = o(N_1)$, $\to 0$ if $N_1 - m = o(N_1)$, and converges to a number on the interval $(0, \tau_1)$ otherwise under Assumption A.2(ii). Note that

$$\begin{aligned}r_{1j}(m) - r_{1j}(N_1) = &\frac{1}{N}\left[\sum_{i=1}^{m} |u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2 - \sum_{i=1}^{N_1} |u_{\pi_j(i),j} - \bar{u}_{1,N_1}(j)|^2\right] \\ &+ \frac{1}{N}\left[\sum_{i=m+1}^{N} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2 - \sum_{i=N_1+1}^{N} |u_{\pi_j(i),j} - \bar{u}_{N_1+1,N}(j)|^2\right] \\ &+ \frac{2a_{1m}}{N} \sum_{i=m+1}^{N_1} [u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] \\ &+ \frac{2}{N}\left\{a_{2m} \sum_{i=N_1+1}^{N_1+N_2} [u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] - a_{2N_1} \sum_{i=N_1+1}^{N_1+N_2} [u_{\pi_j(i),j} - \bar{u}_{N_1+1,N}(j)]\right\} \\ &+ \frac{2}{N}\left\{a_{3m} \sum_{i=N_1+N_2+1}^{N} [u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] - a_{3N_1} \sum_{i=N_1+N_2+1}^{N} [u_{\pi_j(i),j} - \bar{u}_{N_1+1,N}(j)]\right\} \\ \equiv &\, I_{1,m} + I_{2,m} + 2I_{3,m} + 2I_{4,m} + 2I_{5,m}, \text{ say.}\end{aligned}$$

By Theorem 3.1, the fact that $\mu_{1mN} \geq 1/N$ for any $m < N_1$ and that $\bar{u}_{1,N_1}(j) - \bar{u}_{1,m}(j) = \frac{1}{N_1}\sum_{i=1}^{N_1} u_{\pi_j(i),j} - \frac{1}{m}\sum_{i=1}^{m} u_{\pi_j(i),j} = -\frac{N_1-m}{m}[\frac{1}{N_1}\sum_{i=1}^{N_1} u_{\pi_j(i),j} - \frac{1}{N_1-m}\sum_{i=m+1}^{N_1} u_{\pi_j(i),j}]$, we can readily show that uniformly in $m < N_1$

$$\begin{aligned}I_{1,m} &= -\frac{1}{N}\sum_{i=m+1}^{N_1} |u_{\pi_j(i),j}|^2 + \frac{1}{N}\left[N_1|\bar{u}_{1,N_1}(j)|^2 - m|\bar{u}_{1,m}(j)|^2\right] \\ &= -\frac{1}{N}\sum_{i=m+1}^{N_1} |u_{\pi_j(i),j}|^2 + \frac{N_1-m}{N}|\bar{u}_{1,N_1}(j)|^2 + \frac{m}{N}[\bar{u}_{1,N_1}(j) - \bar{u}_{1,m}(j)] \cdot [\bar{u}_{1,N_1}(j) + \bar{u}_{1,m}(j)] \\ &= \frac{N_1-m}{N}\left\{\frac{-1}{N_1-m}\sum_{i=m+1}^{N_1} |u_{\pi_j(i),j}|^2 + |\bar{u}_{1,N_1}(j)|^2 \right. \\ &\qquad \left. - \left(\frac{1}{N_1}\sum_{i=1}^{N_1} u_{\pi_j(i),j} - \frac{1}{N_1-m}\sum_{i=m+1}^{N_1} u_{\pi_j(i),j}\right)[\bar{u}_{1,N_1}(j) + \bar{u}_{1,m}(j)]\right\} \\ &= \mu_{1mN} \cdot O_P(T^{-1}(\ln T)^6) = o_P(\mu_{mN}).\end{aligned}$$

Similarly, noting that $\bar{u}_{N_1+1,N}(j) - \bar{u}_{m+1,N}(j) = \frac{1}{N-N_1}\sum_{i=N_1+1}^{N} u_{\pi_j(i),j} - \frac{1}{N-m}\sum_{i=m+1}^{N} u_{\pi_j(i),j} = \frac{N_1-m}{N-N_1} \times$

$[\frac{1}{N-m}\sum_{i=m+1}^{N} u_{\pi_j(i),j} - \frac{1}{N_1-m}\sum_{i=m+1}^{N_1} u_{\pi_j(i),j}]$, we have

$$
\begin{aligned}
I_{2,m} &= \frac{1}{N}\sum_{i=m+1}^{N_1} |u_{\pi_j(i),j}|^2 + \frac{1}{N}\left[(N-N_1)\,|\bar{u}_{N_1+1,N}(j)|^2 - (N-m)|\bar{u}_{m+1,N}(j)|^2\right] \\
&= \frac{1}{N}\sum_{i=m+1}^{N_1} |u_{\pi_j(i),j}|^2 - \frac{N_1-m}{N}|\bar{u}_{m+1,N}(j)|^2 \\
&\quad + \frac{N-N_1}{N}[\bar{u}_{N_1+1,N}(j) - \bar{u}_{m+1,N}(j)]\cdot[\bar{u}_{N_1+1,N}(j) + \bar{u}_{m+1,N}(j)] \\
&= \frac{N_1-m}{N}\left(\frac{1}{N_1-m}\sum_{i=m+1}^{N_1} |u_{\pi_j(i),j}|^2 - |\bar{u}_{m+1,N}(j)|^2\right) \\
&\quad + \frac{N_1-m}{N}\left(\frac{1}{N-m}\sum_{i=m+1}^{N} u_{\pi_j(i),j} - \frac{1}{N_1-m}\sum_{i=m+1}^{N_1} u_{\pi_j(i),j}\right)[\bar{u}_{N_1+1,N}(j) + \bar{u}_{m+1,N}(j)] \\
&= \mu_{1mN}\cdot O_P(T^{-1}(\ln T)^6) = o_P(\mu_{1mN}),\ \text{and} \\
I_{3,m} &= \frac{N_1-m}{N}\times\frac{a_{1m}}{N_1-m}\sum_{i=m+1}^{N_1}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] \\
&= \mu_{1mN}\cdot O_P(T^{-1/2}(\ln T)^3) = o_P(\mu_{1mN}).
\end{aligned}
$$

For $I_{4,m}$, noting that

$$
\begin{aligned}
|a_{2m} - a_{2N_1}| &= \left|\frac{N_1-m}{N-m}(\alpha^0_{2,j} - \alpha^0_{1,j}) + \left(\frac{N_3}{N-m} - \frac{N_3}{N-N_1}\right)(\alpha^0_{2,j} - \alpha^0_{3,j})\right| \\
&= \frac{N_1-m}{N}\cdot\frac{N}{N-m}\left|(\alpha^0_{2,j} - \alpha^0_{1,j}) + \frac{N_3}{N-N_1}(\alpha^0_{3,j} - \alpha^0_{2,j})\right| = O(\mu_{1mN}) \text{ for all } m < N_1,
\end{aligned}
$$

we have

$$
\begin{aligned}
I_{4,m} &= \frac{a_{2m}}{N}\sum_{i=N_1+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] - \frac{a_{2N_1}}{N}\sum_{i=N_1+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{N_1+1,N}(j)] \\
&= (a_{2m} - a_{2N_1})\,\frac{1}{N}\sum_{i=N_1+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] - \frac{N_2 a_{2N_1}}{N}[\bar{u}_{m+1,N}(j) - \bar{u}_{N_1+1,N}(j)] \\
&= (a_{2m} - a_{2N_1})\,\frac{1}{N}\sum_{i=N_1+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] \\
&\quad + \mu_{1mN}\frac{N_2 a_{2N_1}}{N-N_1}\left[\frac{1}{N-m}\sum_{i=m+1}^{N} u_{\pi_j(i),i} - \frac{1}{N_1-m}\sum_{i=m+1}^{N_1} u_{\pi_j(i),i}\right] \\
&= \mu_{1mN}\cdot O_P(N^{-1/2}(\ln T)^3) + \mu_{1mN}\cdot O_P(N^{-1/2}(\ln T)^3) = o_P(\mu_{1mN}).
\end{aligned}
$$

Similarly, noting that

$$
\begin{aligned}
|a_{3m} - a_{3N_1}| &= \left|\frac{N_1-m}{N-m}(\alpha^0_{3,j} - \alpha^0_{1,j}) + \left(\frac{N_2}{N-m} - \frac{N_2}{N-N_1}\right)(\alpha^0_{3,j} - \alpha^0_{2,j})\right| \\
&= \frac{N_1-m}{N}\frac{N}{N-m}\left|(\alpha^0_{3,j} - \alpha^0_{1,j}) - \frac{N_2}{N-N_1}(\alpha^0_{3,j} - \alpha^0_{2,j})\right| = O(\mu_{1mN}) \text{ for all } m < N_1,
\end{aligned}
$$

we can show that

$$
\begin{aligned}
I_{5,m} &= \frac{a_{3m}}{N}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j)] - \frac{a_{3N_1}}{N}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j}-\bar{u}_{N_1+1,N}(j)] \\
&= (a_{3m}-a_{3N_1})\,\frac{1}{N}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j)] + \frac{N_3 a_{3N_1}}{N}[\bar{u}_{N_1+1,N}(j)-\bar{u}_{m+1,N}(j)] \\
&= \mu_{1mN}\cdot O_P(T^{-1/2}(\ln T)^3) + \mu_{1mN}\frac{N_3 a_{3N_1}}{N-N_1}\left[\frac{1}{N-m}\sum_{i=m+1}^{N}u_{\pi_j(i),j} - \frac{1}{N_1-m}\sum_{i=m+1}^{N_1}u_{\pi_j(i),j}\right] \\
&= \mu_{1mN}\cdot O_P(T^{-1/2}(\ln T)^3) + \mu_{1mN}\cdot O_P(T^{-1/2}(\ln T)^3) = o_P(\mu_{1mN}).
\end{aligned}
$$

Thus $r_{1j}(m)-r_{1j}(N_1)=o_P(\mu_{1mN})$ uniformly in $m<N_1$ and

$$
\begin{aligned}
P(\hat{m}_1(j)<N_1) &\le P(\exists m<N_1, S_{1,N}(j,m)-S_{1,N}(j,N_1)<0) \\
&= P(\exists m<N_1, \Delta M_1(m)+[r_1(m)-r_1(N_1)]<0)\to 0
\end{aligned}
$$

as $(N,T)\to\infty$. Then (i) follows.

We now study case (ii). Noting that when $N_1<m\le N_1+N_2$, $\bar{\beta}_{1,m}(j)=\frac{N_1}{m}\alpha^0_{1,j}+\frac{m-N_1}{m}\alpha^0_{2,j}+\bar{u}_{1,m}(j)$, and $\bar{\beta}_{m+1,N}(j)=\frac{N_1+N_2-m}{N-m}\alpha^0_{2,j}+\frac{N_3}{N-m}\alpha^0_{3,j}+\bar{u}_{m+1,N}(j)$. It follows that for $i=1,\dots,m$,

$$
\tilde{\beta}_{\pi_j(i),j}-\bar{\beta}_{1,m}(j)=\begin{cases} b_{1m}+u_{\pi_j(i),j}-\bar{u}_{1,m}(j) & \text{if } \pi_j(i)\in G_1^0,\\ b_{2m}+u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j) & \text{if } \pi_j(i)\in G_2^0,\end{cases}
$$

where $b_{1m}=\frac{m-N_1}{m}(\alpha^0_{1,j}-\alpha^0_{2,j})$ and $b_{2m}=\frac{N_1}{m}(\alpha^0_{2,j}-\alpha^0_{1,j})$. So for the left segment, we have

$$
\begin{aligned}
\sum_{i=1}^{m}|\tilde{\beta}_{\pi_j(i),j}-\bar{\beta}_{1,m}(j)|^2 &= \sum_{i=1}^{N_1}|\tilde{\beta}_{\pi_j(i),j}-\bar{\beta}_{1,m}(j)|^2 + \sum_{i=N_1+1}^{m}|\tilde{\beta}_{\pi_j(i),j}-\bar{\beta}_{1,m}(j)|^2 \\
&= N_1|b_{1m}|^2 + (m-N_1)|b_{2m}|^2 + 2b_{1m}\sum_{i=1}^{N_1}[u_{\pi_j(i),j}-\bar{u}_{1,m}(j)] \\
&\quad + 2b_{2m}\sum_{i=N_1+1}^{m}[u_{\pi_j(i),j}-\bar{u}_{1,m}(j)] + \sum_{i=1}^{m}|u_{\pi_j(i),j}-\bar{u}_{1,m}(j)|^2.
\end{aligned}
$$

Similarly for $i=m+1,\dots,N$, we have

$$
\tilde{\beta}_{\pi_j(i),j}-\bar{\beta}_{m+1,N}(j)=\begin{cases} b_{3m}+u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j) & \text{if } \pi_j(i)\in G_2^0,\\ b_{4m}+u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j) & \text{if } \pi_j(i)\in G_3^0,\end{cases}
$$

where $b_{3m}=\frac{N_3}{N-m}(\alpha^0_{2,j}-\alpha^0_{3,j})$ and $b_{4m}=\frac{N_1+N_2-m}{N-m}(\alpha^0_{3,j}-\alpha^0_{2,j})$. Note that b_{1m}, b_{2m}, b_{3m}, and b_{4m} are each $O(1)$ uniformly in $m\in\{N_1+1,\dots,N_1+N_2\}$. Then for the right segment, we get

$$
\begin{aligned}
\sum_{i=m+1}^{N}|\tilde{\beta}_{\pi_j(i),j}-\bar{\beta}_{m+1,N}(j)|^2 &= (N_1+N_2-m)|b_{3m}|^2 + N_3|b_{4m}|^2 + 2b_{3m}\sum_{i=m+1}^{N_1+N_2}[u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j)] \\
&\quad + 2b_{4m}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j)] + \sum_{i=m+1}^{N}[u_{\pi_j(i),j}-\bar{u}_{m+1,N}(j)]^2.
\end{aligned}
$$

Combining the above expressions yields $S_{1,N}(j,m) = M_{2j}(m) + r_{2j}(m)$, where $M_{2j}(m) = \frac{N_1}{N}|b_{1m}|^2 + \frac{m-N_1}{N}|b_{2m}|^2 + \frac{N_1+N_2-m}{N}|b_{3m}|^2 + \frac{N_3}{N}|b_{4m}|^2$, and

$$\begin{aligned} r_{2j}(m) = & \frac{1}{N}\sum_{i=1}^{m}|u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2 + \frac{1}{N}\sum_{i=m+1}^{N}|u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2 \\ & + \frac{2b_{1m}}{N}\sum_{i=1}^{N_1}[u_{\pi_j(i),j} - \bar{u}_{1,m}(j)] + \frac{2b_{2m}}{N}\sum_{i=N_1+1}^{m}[u_{\pi_j(i),j} - \bar{u}_{1,m}(j)] \\ & + \frac{2b_{3m}}{N}\sum_{i=m+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] + \frac{2b_{4m}}{N}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)]. \end{aligned}$$

Now for $m \in \{N_1+1,\ldots,N_1+N_2\}$, we have $S_{1,N}(j,m) - S_{1,N}(j,N_1+N_2) = [M_{2j}(m) - M_{2j}(N_1+N_2)] + [r_{2j}(m) - r_{2j}(N_1+N_2)]$, where a rough uniform bound for $r_{2j}(m) - r_{2j}(N_1)$ is $O_P(T^{-1/2}(\ln T)^3)$ by Theorem 3.1. Note that $b_{1N_1} = 0$, we have

$$\begin{aligned} \Delta M_{2j}(m) \equiv & M_{2j}(m) - M_{2j}(N_1) \\ = & \frac{N_1}{N}|b_{1m}|^2 + \frac{m-N_1}{N}|b_{2m}|^2 + \frac{N_1+N_2-m}{N}|b_{3m}|^2 + \frac{N_3}{N}|b_{4m}|^2 - \frac{N_2}{N}|b_{3N_1}|^2 - \frac{N_3}{N}|b_{4N_1}|^2 \\ = & \frac{m-N_1}{N}\left[\frac{N_1}{m}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{N_3^2}{(N-N_1)(N-m)}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2\right] \\ \geq & \frac{m-N_1}{N}\frac{N_1+N_2}{m}\left[\frac{N_1}{N_1+N_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{N_3}{N-N_1}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2\right] \\ = & \mu_{2mN}\frac{N_1+N_2}{m}\left[\frac{\tau_1}{\tau_1+\tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{\tau_3}{\tau_2+\tau_3}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2\right]\cdot[1+o(1)] \\ \asymp & \mu_{2mN} \text{ uniformly in } m \in \{N_1+1,\ldots,N_1+N_2\}, \end{aligned}$$

where the inequality follows from the fact that $\frac{N_3}{N-m} \leq 1 \leq \frac{N_1+N_2}{m}$ for all $m \in \{N_1+1,\ldots,N_1+N_2\}$, $\mu_{2mN} = \frac{m-N_1}{N}$, and the last line follows from the fact that $\frac{\tau_1}{\tau_1+\tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{\tau_3}{\tau_2+\tau_3}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2 > 0$ in Case 1a. Following the analysis of $r_{1j}(m) - r_{1j}(N_1)$, we can show that $r_{2j}(m) - r_{2j}(N_1) = o_P(\mu_{2mN})$ uniformly in $m \in \{N_1+1,\ldots,N_1+N_2\}$. It follows that as $(N,T) \to \infty$, for any $j \in \mathcal{S}_2$ we must have

$$P(N_1 < \hat{m}_1 \leq N_1+N_2) = P(\exists m \in \{N_1+1,\ldots,N_1+N_2\} \text{ s.t. } S_{1,N}(j,m) - S_{1,N}(j,N_1) < 0) \to 0.$$

Analogously, we can show (iii). It follows that $P(\hat{m}_1(j) = N_1) \to 1$ as $(N,T) \to \infty$ in Case 1a. In other words, by using the ranking relation (B.1) based on regressor $\hat{j}_1 = j \in \mathcal{S}_1$, we could find the right break point w.p.a.1. in the first round of the SBSA. For the ease of presentation, we continue to use $j \in \mathcal{S}_1$ to represent $\hat{j}_1$. Given the first identified break point being $\{N_1\}$ in Case 1a, we have

$$\begin{aligned} S_{1,N}(j,N_1) = & \frac{1}{N}\left[\sum_{1\leq i\leq N_1}\left|\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{1,N_1}(j)\right|^2 + \sum_{N_1+1\leq i\leq N}\left|\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{N_1+1,N}(j)\right|^2\right] \\ = & \frac{1}{N}\left[N_2\left|\alpha_{2,j}^0 - \frac{N_2}{N_2+N_3}\alpha_{2,j}^0 - \frac{N_3}{N_2+N_3}\alpha_{3,j}^0\right|^2 + N_3\left|\alpha_{3,j}^0 - \frac{N_2}{N_2+N_3}\alpha_{2,j}^0 - \frac{N_3}{N_2+N_3}\alpha_{3,j}^0\right|^2\right] \\ & + o_P(1) \\ = & \frac{1}{N}\frac{N_2N_3}{N_2+N_3}\left|\alpha_{2,j}^0 - \alpha_{3,j}^0\right|^2 + o_P(1) \xrightarrow{P} \frac{\tau_2\tau_3}{\tau_2+\tau_3}c_{23,j}^0 \equiv \Delta_{1,j}, \end{aligned}$$

where $c_{23,j}^0 = |\alpha_{2,j}^0 - \alpha_{3,j}^0|^2$. Now we have two segmentations with one containing elements in Group 1 and the other containing elements in Groups 2 and 3 w.p.a.1. That is,

$$P\left(\hat{G}_1(2) = \{\pi_j(1),\ldots,\pi_j(N_1)\},\ \hat{G}_2(2) = \{\pi_j(N_1+1),\ldots,\pi_j(N)\}\right) \to 1.$$

In the second step, we repeat the iterative algorithm on $\hat{G}_1(2)$ and $\hat{G}_2(2)$. When $\hat{\jmath}_1 \in \mathcal{S}_1$ and $\{N_1\}$ is identified in the first step w.p.a.1, $\hat{\jmath}_2$ can belong to either $\mathcal{S}_1$ or $\mathcal{S}_2$. We show that no matter which value $\hat{\jmath}_2$ takes in $\mathcal{S}_1 \cup \mathcal{S}_2$, we can identify the second break point $\{N_1 + N_2\}$ w.p.a.1. For the binary segments over $\hat{G}_2(2)$, following similar arguments leading to $P(\hat{m}_1(j) = N_1) \to 1$ as $(N,T) \to \infty$, we can show that $P(\hat{m}_2(j) = N_2) \to 1$ for any $j \in \mathcal{S}_1 \cup \mathcal{S}_2$ (and thus also for $\hat{\jmath}_2$) where $\hat{m}_2(j) = \arg\min_{1 \le m < N_2+N_3} S_{N_1+1,N}(j,m)$. Then $\hat{G}_2(2)$ is divided into two sub-segments $\hat{G}_{21}(2)$ and $\hat{G}_{22}(2)$ such that

$$P\left(\hat{G}_{21}(2) = \{\pi_j(N_1+1), \ldots, \pi_j(N_1+N_2)\},\ \hat{G}_{22}(2) = \{\pi_j(N_1+N_2+1), \ldots, \pi_j(N)\}\right) \to 1.$$

Furthermore, we can show that by using arguments used in the proof of Theorem 3.1

$$\frac{1}{N}\left(\sum_{i\in\hat{G}_1(2)}\left|\tilde{\beta}_{i,j} - \bar{\beta}_{\hat{G}_1(2)}(j)\right|^2 + \sum_{i\in\hat{G}_{21}(2)}\left|\tilde{\beta}_{i,j} - \bar{\beta}_{\hat{G}_{21}(2)}(j)\right|^2 + \sum_{i\in\hat{G}_{22}(2)}\left|\tilde{\beta}_{i,j} - \bar{\beta}_{\hat{G}_{22}(2)}(j)\right|^2\right) = O_P(T^{-1}),$$

where $\bar{\beta}_{\hat{G}_1(2)}(j) = \frac{1}{|\hat{G}_1(2)|}\sum_{i\in\hat{G}_1(2)}\tilde{\beta}_{i,j}$, and $\bar{\beta}_{\hat{G}_{21}(2)}(j)$ and $\bar{\beta}_{\hat{G}_{22}(2)}(j)$ are similarly defined. In contrast, for any binary segments $\{\hat{G}_{11}(2), \hat{G}_{12}(2)\}$ over $\hat{G}_1(2)$, we can show that

$$\frac{1}{N}\left(\sum_{i\in\hat{G}_{11}(2)}\left|\tilde{\beta}_{i,j} - \bar{\beta}_{\hat{G}_{11}(2)}(j)\right|^2 + \sum_{i\in\hat{G}_{12}(2)}\left|\tilde{\beta}_{i,j} - \bar{\beta}_{\hat{G}_{12}(2)}(j)\right|^2 + \sum_{i\in\hat{G}_2(2)}\left|\tilde{\beta}_{i,j} - \bar{\beta}_{\hat{G}_2(2)}(j)\right|^2\right) = \Delta_{1,j} + o_P(1),$$

because

$$\begin{aligned}
\Delta_{1,j} &= \lim_{N\to\infty}\frac{1}{N}\sum_{i=N_1+1}^{N}\left|\beta^0_{\pi_j(i),j} - \frac{1}{N-N_1}\sum_{i'=N_1+1}^{N}\beta^0_{\pi_j(i'),j}\right|^2 \\
&= \lim_{N\to\infty}\frac{1}{N}\left(N_2\left|\alpha^0_{2,j} - \frac{N_2}{N_2+N_3}\alpha^0_{2,j} - \frac{N_3}{N_2+N_3}\alpha^0_{3,j}\right|^2 + N_3\left|\alpha^0_{3,j} - \frac{N_2}{N_2+N_3}\alpha^0_{2,j} - \frac{N_3}{N_2+N_3}\alpha^0_{3,j}\right|^2\right) \\
&= \lim_{N\to\infty}\frac{1}{N}\frac{N_2N_3}{N_2+N_3}\left|\alpha^0_{2,j} - \alpha^0_{3,j}\right|^2 = \frac{\tau_2\tau_3}{\tau_2+\tau_3}c^0_{23,j} > 0.
\end{aligned}$$

Then based on our SBSA, $N_1 + N_2$ will be identified as the second break point.

In sum, if $\hat{\jmath}_1 \in \mathcal{S}_1$ and Case 1a is in effect, we have shown $\{N_1\}$ is identified in the first step, and no matter what value $\hat{\jmath}_2$ takes in $\mathcal{S}_1 \cup \mathcal{S}_2$, we can identify the second break point $\{N_1 + N_2\}$ in the second step. This results 3 groups with $\hat{\mathcal{G}}(3) = \{\hat{G}_1(2), \hat{G}_{21}(2), \hat{G}_{22}(2)\}$ such that $P(\hat{\mathcal{G}}(3) = \mathcal{G}^0) \to 1$ as $(N,T) \to \infty$.

Analogously, when $\hat{\jmath}_1 \in \mathcal{S}_1$ and Case 1b is in effect, we can show that $\{N_1 + N_2\}$ is identified in the first step and as mentioned above, in the second step our algorithm ensures that $\hat{\jmath}_2 \in \mathcal{S}_1$ and $\{N_1\}$ is identified w.p.a.1. When $\hat{\jmath}_1 \in \mathcal{S}_1$ and Case 1c is in effect, we can follow Bai (1997) and show that each of $\{N_1\}$ and $\{N_1 + N_2\}$ can be identified in the first step with probability $\frac{1}{2}$, and the other point will be identified in the second step w.p.a.1.

Since $K^0 = 3$ is known, the algorithm stops here and the proof in Case 1 is completed.

Case 2: $\hat{\jmath}_1 \in \mathcal{S}_2$. W.l.o.g., we consider $\alpha^0_{1,j} = \alpha^0_{2,j} < \alpha^0_{3,j}$ for $\hat{\jmath}_1 = j \in \mathcal{S}_2$. In this case, it is impossible to distinguish elements from Group 1 from those from Group 2 according to the regressor-j-based ranking relation in (B.1). Now based on (B.1) and the fact that $\max_{1\le i\le N}\|u_i\| = O_P(T^{-1/2}(\ln T)^3)$, we have the following homogeneity property

$$\beta^0_{\pi_j(i),j} = \begin{cases} \alpha^0_{1,j} = \alpha^0_{2,j} & 1 \le i \le N_1 + N_2, \\ \alpha^0_{3,j} & N_1 + N_2 + 1 \le i \le N. \end{cases}$$

Recall that $\hat{m}_1 = \arg\min_{1\leq m<N} S_{1,N}(j,m)$ where we suppress the dependence of $\hat{m}_1$ on j. As in Case 1, we want to show as $(N,T)\to\infty$, $P(\hat{m}_1 = N_1+N_2)\to 1$ for any $j\in\mathcal{S}_2$ by showing that (i1) $P(\hat{m}_1 < N_1+N_2)\to 0$ and (i2) $P(\hat{m}_1 > N_1+N_2)\to 0$.

First, we consider the case where $m < N_1+N_2$. Note that $\bar{\beta}_{1,m}(j) = \alpha^0_{1,j} + \bar{u}_{1,m}(j)$ and $\bar{\beta}_{m+1,N}(j) = \frac{N_1+N_2-m}{N-m}\alpha^0_{1,j} + \frac{N_3}{N-m}\alpha^0_{3,j} + \bar{u}_{m+1,N}(j)$. It follows that for $i=1,\ldots,m$, $\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{1,m}(j) = u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)$, and

$$\sum_{i=1}^{m}|\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{1,m}(j)|^2 = \sum_{i=1}^{m}|u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2.$$

Similarly for $i = m+1,\ldots,N$, we have

$$\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{m+1,N}(j) = \begin{cases} c_{1m} + u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j) & \text{if } \pi_j(i)\in G^0_1\cup G^0_2, \\ c_{2m} + u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j) & \text{if } \pi_j(i)\in G^0_3, \end{cases}$$

where $c_{1m} = \frac{N_3}{N-m}(\alpha^0_{1,j} - \alpha^0_{3,j})$ and $c_{2m} = \frac{N_1+N_2-m}{N-m}(\alpha^0_{3,j} - \alpha^0_{1,j})$. Note that c_{1m} and c_{2m} are each $O(1)$ uniformly in $m < N_1+N_2$. Then

$$\begin{aligned}\sum_{i=m+1}^{N}|\tilde{\beta}_{\pi_j(i),j} - \bar{\beta}_{m+1,N}(j)|^2 &= (N_1+N_2-m)|c_{1m}|^2 + N_3|c_{2m}|^2 + 2c_{1m}\sum_{i=m+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] \\ &\quad + 2c_{2m}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] + \sum_{i=m+1}^{N}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)]^2.\end{aligned}$$

Combining the expressions above yields $S_{1,N}(j,m) = M_{3j}(m) + r_{3j}(m)$, where $M_{3j}(m) = \frac{N_1+N_2-m}{N}|c_{1m}|^2 + \frac{N_3}{N}|c_{2m}|^2$, and

$$\begin{aligned}r_{3j}(m) &= \frac{1}{N}\sum_{i=1}^{m}|u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2 + \frac{1}{N}\sum_{i=m+1}^{N}|u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2 \\ &\quad + \frac{2c_{1m}}{N}\sum_{i=m+1}^{N_1+N_2}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)] + \frac{2c_{2m}}{N}\sum_{i=N_1+N_2+1}^{N}[u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)].\end{aligned}$$

Now for $m < N_1+N_2$, we have $S_{1,N}(j,m) - S_{1,N}(j,N_1+N_2) = [M_{3j}(m) - M_{3j}(N_1+N_2)] + [r_{3j}(m) - r_{3j}(N_1+N_2)]$. Noting that $c_{2,N_1+N_2} = 0$, we have

$$\begin{aligned}\Delta M_{3j}(m) \equiv M_{3j}(m) - M_{3j}(N_1+N_2) &= \frac{N_1+N_2-m}{N}|c_{1m}|^2 + \frac{N_3}{N}|c_{2m}|^2 \\ &= \left[\frac{N_1+N_2-m}{N}\left(\frac{N_3}{N-m}\right)^2 + \frac{N_3}{N}\left(\frac{N_1+N_2-m}{N-m}\right)^2\right](\alpha^0_{1,j} - \alpha^0_{3,j})^2 \\ &= \frac{N_3(N_1+N_2-m)}{N(N-m)}(\alpha^0_{1,j} - \alpha^0_{3,j})^2 \asymp \mu_{3mN},\end{aligned}$$

where $\mu_{3mN} = (N_1+N_2-m)/N$. Following the analysis of $r_{1j}(m) - r_{1j}(N_1)$, we can show that $r_{3j}(m) - r_{3j}(N_1+N_2) = o_P(\mu_{3mN})$ uniformly in $m < N_1+N_2$. It follows that as $(N,T)\to\infty$, for any $j\in\mathcal{S}_2$ we must have

$$P(\hat{m}_1 < N_1+N_2) = P(\exists m < N_1+N_2 \text{ s.t. } S_{1,N}(j,m) - S_{1,N}(j,N_1+N_2) < 0)\to 0. \tag{B.2}$$

By mere symmetry, we can prove that as $(N,T)\to\infty$,

$$P(\hat{m}_1 > N_1+N_2) = P(\exists m > N_1+N_2 \text{ s.t. } S_{1,N}(j,m) - S_{1,N}(j,N_1+N_2) < 0)\to 0. \tag{B.3}$$

Combining (B.2) and (B.3), we have as $(N,T)\to\infty$, $P(\hat{m}_1(j)=N_1+N_2)\to 1$ and

$$P\left(\hat{G}_1(2)=\{\pi_j(1),\ldots,\pi_j(N_1+N_2)\} \text{ and } \hat{G}_2(2)=\{\pi_j(N_1+N_2+1),\ldots,\pi_j(N)\}\right)\to 1.$$

Recall here regressor j is a representative element in set $\mathcal{S}_2$.

The above proof applies when $\hat{j}_1\in\mathcal{S}_2$. In the second step, our algorithm ensures $\hat{j}_2\in\mathcal{S}_1$ because the segment $\mathcal{S}_{1,N_1+N_2}(j)$ contains no break point for any $j\in\mathcal{S}_2$. Following the analysis in Case 1a, we can readily show that for any $j\in\mathcal{S}_1$, we can identify the second break point $\{N_1\}$ in the second step through our SBSA. As a result, we have $P(\hat{m}_1=N_1+N_2,\ \hat{m}_2=N_1)\to 1$ and $P(\hat{\mathcal{G}}(3)=\mathcal{G}^0)\to 1$ as $(N,T)\to\infty$. That is, we can identify all three groups w.p.a.1. This completes the proof of the theorem for the case $K^0=3$. When $K^0>3$, we need to deal with extra terms. But by similar arguments as that of Bai (1997) and KLZ, the proof strategy is essentially the same and the two break points case doesn't lose generality. ■

Proof of Theorem 3.3: By Theorem 3.2 and Assumption A3,

$$\begin{aligned}2Q_{NT}(\hat{\boldsymbol{\beta}}(K^0),\hat{\boldsymbol{\theta}}(K^0)) &= 2Q_{NT}(\hat{\boldsymbol{\beta}}(K^0),\hat{\boldsymbol{\theta}}(K^0))[\mathbf{1}\{\hat{\mathcal{G}}\left(K^0\right)=\mathcal{G}^0\}+\mathbf{1}\{\hat{\mathcal{G}}\left(K^0\right)\neq\mathcal{G}^0\}]\\ &= 2Q_{NT}(\hat{\boldsymbol{\beta}}(K^0),\hat{\boldsymbol{\theta}}(K^0))\mathbf{1}\{\hat{\mathcal{G}}\left(K^0\right)=\mathcal{G}^0\}+o_P(1)\\ &\to\sigma_0^2 \text{ as } (N,T)\to\infty.\end{aligned}$$

Then $\mathrm{IC}_1(K^0)=2Q_{NT}(\hat{\boldsymbol{\beta}}(K^0),\hat{\boldsymbol{\theta}}(K^0))+pK^0\cdot\rho_{NT}\to\sigma_0^2$ by Assumption A3(iii).

When $1\le K<K^0$, by Assumption A3(ii) we have

$$\begin{aligned}\mathrm{IC}_1(K) &= 2Q_{NT}(\hat{\boldsymbol{\beta}}(K),\hat{\boldsymbol{\theta}}(K))+pK\cdot\rho_{NT}\ge 2\min_{1\le K<K^0}\min_{\mathcal{G}(K)}\hat{\sigma}^2_{\mathcal{G}(K)}+pK\cdot\rho_{NT}\\ &\to\bar{\sigma}^2>\sigma_0^2 \text{ as } (N,T)\to\infty.\end{aligned}$$

So we have

$$P(\hat{K}<K^0)=P(\exists 1\le K<K^0,\ \mathrm{IC}_1(K)<\mathrm{IC}_1(K^0))\to 0 \text{ as } (N,T)\to\infty. \tag{B.4}$$

Next, we consider the case where $K^0<K\le K^{\max}$. By Theorem 3.2, the true group structure will be identified w.p.a.1 when K^0 is known. When $K>K^0$, so we get a further unnecessary refinement of the true group structure. Following the analysis of Lemma S1.14 in SSPb, we can readily show that $T\max_{K^0<K\le K^{\max}}(\hat{\sigma}^2_{\hat{\mathcal{G}}(K)}-\hat{\sigma}^2_{\hat{\mathcal{G}}(K^0)})=O_P(1)$. It follows that by Assumption A3(iii)

$$\begin{aligned}P(\hat{K}>K^0) &= P(\exists K^0<K\le K^{\max},\ \mathrm{IC}_1(K)<\mathrm{IC}_1(K^0))\\ &= P(\exists K^0<K\le K^{\max}, T(\hat{\sigma}^2_{\hat{\mathcal{G}}(K)}-\hat{\sigma}^2_{\hat{\mathcal{G}}(K^0)})>(K-K^0)T\rho_{NT})\\ &\to 0 \text{ as } (N,T)\to\infty.\end{aligned} \tag{B.5}$$

Combining (B.4) and (B.5), we have $P(\hat{K}=K^0)\to 1$ as $(N,T)\to\infty$. ■

Proof of Theorem 3.4: Let $D_{NT}=\mathrm{diag}(\sqrt{N_1T}I_p,\ldots,\sqrt{N_{K^0}T}I_p,\sqrt{NT}I_q)$, $E_{NT}=\{\hat{\mathcal{G}}(\hat{K})=\mathcal{G}^0\}$, $\Xi_{NT}=D_{NT}((\hat{\alpha}_1-\alpha_1^0)^\top,\ldots,(\hat{\alpha}_{K^0}-\alpha_{K^0}^0)^\top,(\hat{\theta}-\theta^0)^\top)^\top$, and $\Xi^*_{NT}=D_{NT}((\hat{\alpha}_1^*-\alpha_1^0)^\top,\ldots,(\hat{\alpha}^*_{K^0}-\alpha^0_{K^0})^\top,(\hat{\theta}^*-\theta^0)^\top)^\top$. Then $P(E_{NT})\to 1$ as $(N,T)\to\infty$ by Theorems 3.2 and 3.3 and

$$\begin{aligned}P(\Xi_{NT}\in\mathcal{C}) &= P(\Xi_{NT}\in\mathcal{C} \text{ and } E_{NT})+P(\Xi_{NT}\in\mathcal{C} \text{ and } E^c_{NT})\\ &= P(\Xi^*_{NT}\in\mathcal{C})+o(1),\end{aligned}$$

where E^c_{NT} denote the complement of E_{NT} and $\mathcal{C}\subset\mathbb{R}^{K^0p+q}$. That is, it suffices to consider the asymptotic distribution of the oracle estimators $\hat{\alpha}_1^*,\ldots,\hat{\alpha}^*_{K^0}$, and $\hat{\theta}^*$.

Consider the minimization of the profile log-likelihood function in (2.10) with $\hat{\mathcal{G}}(\hat{K})$ being replaced by $\mathcal{G}^0$. By the envelope theorem, the first order conditions with respect to α_k and θ are respectively by

$$\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^{T} U(w_{it};\hat{\alpha}_k^*,\hat{\mu}_i(\hat{\alpha}_k^*,\hat{\theta}^*),\hat{\theta}^*) = \mathbf{0}_{p\times 1} \text{ for } k=1,\ldots,K^0, \text{ and} \tag{B.6}$$

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W(w_{it};\hat{\alpha}_k^*,\hat{\mu}_i(\hat{\alpha}_k^*,\hat{\theta}^*),\hat{\theta}^*) = \mathbf{0}_{q\times 1}. \tag{B.7}$$

By Taylor expansions, we have

$$-\begin{bmatrix} \breve{H}_{11} & \cdots & 0 & \breve{H}_{1\theta} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \breve{H}_{K^0K^0} & \breve{H}_{K^0\theta} \\ \breve{H}_{\theta 1} & \cdots & \breve{H}_{\theta K^0} & \breve{H}_{\theta\theta} \end{bmatrix}\begin{bmatrix} \hat{\alpha}_1^* - \alpha_1^0 \\ \vdots \\ \hat{\alpha}_{K^0}^* - \alpha_{K^0}^0 \\ \hat{\theta}^* - \theta^0 \end{bmatrix} = \begin{bmatrix} \frac{1}{N_k T}\sum_{i\in G_1^0}\sum_{t=1}^{T} U(w_{it};\alpha_1^0,\hat{\mu}_i(\alpha_1^0,\theta^0),\theta^0) \\ \vdots \\ \frac{1}{N_k T}\sum_{i\in G_{K^0}^0}\sum_{t=1}^{T} U(w_{it};\alpha_{K^0}^0,\hat{\mu}_i(\alpha_{K^0}^0,\theta^0),\theta^0) \\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0,\theta^0),\theta^0) \end{bmatrix}, \tag{B.8}$$

where for $k=1,\ldots,K^0$,

$$\breve{H}_{kk} \equiv \frac{1}{N_k}\sum_{i\in G_k^0}\hat{H}_{i,\beta\beta}(\breve{\alpha}_k,\breve{\theta}), \qquad \breve{H}_{k\theta} \equiv \frac{1}{N_k}\sum_{i\in G_k^0}\hat{H}_{i,\beta\theta}(\breve{\alpha}_k,\breve{\theta}),$$

$$\breve{H}_{\theta k} \equiv \frac{1}{N}\sum_{i=1}^{N}\hat{H}_{i,\theta\beta}(\breve{\alpha}_k,\breve{\theta}), \qquad \breve{H}_{\theta\theta} \equiv \frac{1}{N}\sum_{i=1}^{N}\hat{H}_{i,\theta\theta}(\breve{\alpha}_k,\breve{\theta}),$$

$\hat{H}_{i,\beta\beta}(\alpha_k,\theta)$, $\hat{H}_{i,\beta\theta}(\alpha_k,\theta)$, $\hat{H}_{i,\theta\beta}(\alpha_k,\theta)$, and $\hat{H}_{i,\theta\theta}(\alpha_k,\theta)$ are defined analogously to $H_{i,\beta\beta}(\beta_i,\theta)$, $H_{i,\beta\theta}(\beta_i,\theta)$, $H_{i,\theta\beta}(\beta_i,\theta)$, and $H_{i,\theta\theta}(\beta_i,\theta)$ below (3.2) with $\mu_i(\beta_i,\theta)$ being replaced by $\hat{\mu}_i(\alpha_k,\theta)$, $\breve{\alpha}_k$ lies between $\hat{\alpha}_k^*$ and α_k^0 elementwise, and $\breve{\theta}$ lies between $\hat{\theta}^*$ and θ^0 elementwise. Following the analysis of Theorem 3.1, we can show the consistency of $\hat{\alpha}_k^*$ and $\hat{\theta}^*$. With this result, we can follow the proof of Lemma S1.13 in SSPb (see also the proof of Lemma A.2) and show that

$$\breve{H}_{kk} = H_{kk} + o_P(1), \quad \breve{H}_{k\theta} = H_{k\theta} + o_P(1), \quad \breve{H}_{\theta k} = H_{\theta k} + o_P(1), \text{ and } \breve{H}_{\theta\theta} = H_{\theta\theta} + o_P(1),$$

where, e.g., $H_{kk} = \frac{1}{N_k}\sum_{i\in G_k^0}\mathbb{E}\left[H_{i,\beta\beta}\left(\beta_i^0,\theta^0\right)\right]$, $H_{k\theta}$, $H_{\theta k}$, and $H_{\theta\theta}$ are analogously defined.

Let $\mathbb{S}_{kNT} = \frac{1}{\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T} U(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0,\theta^0),\theta^0)$ for $k=1,\ldots,K^0$ and $\mathbb{S}_{\theta NT} = \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} W(w_{it};\beta_i^0,\hat{\mu}_i(\beta_i^0,\theta^0),\theta^0)$. As in the proof of Lemma S1.12 in SSPb, we apply the second order Taylor expansion to obtain

$$\begin{aligned}\mathbb{S}_{kNT} &= \frac{1}{\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T} U_{it} + \frac{1}{\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T} U_{it}^{\mu}[\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0] + \frac{1}{2\sqrt{N_k T}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\hat{s}_{it,U} \\ &\equiv \mathbb{S}_{kNT,1} + \mathbb{S}_{kNT,2} + \mathbb{S}_{kNT,3}, \end{aligned} \tag{B.9}$$

$$\begin{aligned}\mathbb{S}_{\theta NT} &= \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} + \frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}^{\mu}[\hat{\mu}_i(\beta_i^0,\theta^0)-\mu_i^0] + \frac{1}{2\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{s}_{it,W} \\ &\equiv \mathbb{S}_{\theta NT,1} + \mathbb{S}_{\theta NT,2} + \mathbb{S}_{\theta NT,3}, \end{aligned} \tag{B.10}$$

where $[\hat{s}_{it,U}]_j \equiv [\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0]^\top U_j^{\mu\mu}(w_{it};\alpha_k^0,\breve{\mu}_i,\theta^0)[\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0]$, $[\hat{s}_{it,W}]_j \equiv [\hat{\mu}_i(\beta_i^0,\theta^0)-\mu_i^0]^\top W_j^{\mu\mu}(w_{it};\beta_i^0,\breve{\mu}_i,\theta^0)[\hat{\mu}_i(\beta_i^0,\theta^0)-\mu_i^0]$, $U_j^{\mu\mu}(w_{it};\alpha_k,\mu_i,\theta)$ denotes the second order partial derivatives of the jth element of $U(w_{it};\alpha_k,\mu_i,\theta)$ with respect to μ_i, $W_j^{\mu\mu}(w_{it};\beta_i,\mu_i,\theta)$ is similarly defined, and both $\breve{\mu}_i$ and $\breve{\mu}_i$ lie between $\hat{\mu}_i(\beta_i^0,\theta^0)$ and μ_i^0 elementwise.

To study $\mathbb{S}_{kNT,3}$ and $\mathbb{S}_{\theta NT,3}$, we consider the first order Taylor expansion:

$$\begin{aligned}0 &= \frac{1}{T}\sum_{t=1}^{T} V\left(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0,\theta^0),\theta^0\right)\\ &= \frac{1}{T}\sum_{t=1}^{T} V_{it} + \frac{1}{T}\sum_{t=1}^{T} V^{\mu}\left(w_{it};\alpha_k^0,\bar{\mu}_i(\alpha_k^0,\theta^0),\theta^0\right)\left[\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0\right],\end{aligned}$$

where $\bar{\mu}_i(\alpha_k^0,\theta^0)$ lies between $\hat{\mu}_i(\alpha_k^0,\theta^0)$ and μ_i^0. Solving for $\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0$ and following the proof of Lemma S1.12 in SSPb, we can show that

$$\begin{aligned}\left[\mathbb{S}_{kNT,3}\right]_j &= \frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^{\top}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}U_{it,j}^{\mu\mu}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)+o_P(1)\\ &= \frac{1}{2\sqrt{N_kT}}\sum_{i\in G_k^0}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)^{\top}S_{iV}^{-1}S_{iU2,j}S_{iV}^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)+o_P(1), \qquad \text{(B.11)}\\ \left[\mathbb{S}_{\theta NT,3}\right]_j &= \frac{1}{2\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)^{\top}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}W_{it,j}^{\mu\mu}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}\right)+o_P(1)\\ &= \frac{1}{2\sqrt{NT}}\sum_{i=1}^{N}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)^{\top}S_{iV}^{-1}S_{iW2,j}S_{iV}^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{it}\right)+o_P(1). \qquad \text{(B.12)}\end{aligned}$$

To study $\mathbb{S}_{kNT,2}$ and $\mathbb{S}_{\theta NT,2}$, we need to consider the second order Taylor expansion:

$$\begin{aligned}0 &= \frac{1}{T}\sum_{t=1}^{T} V\left(w_{it};\alpha_k^0,\hat{\mu}_i(\alpha_k^0,\theta^0),\theta^0\right)\\ &= \frac{1}{T}\sum_{t=1}^{T} V_{it} + \frac{1}{T}\sum_{t=1}^{T} V_{it}^{\mu}\left[\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0\right] + \frac{1}{2T}\sum_{t=1}^{T}\hat{s}_{it,V},\end{aligned}$$

where $\left[\hat{s}_{it,V}\right]_j \equiv \left[\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0\right]^{\top}V_j^{\mu\mu}\left(w_{it};\alpha_k^0,\vec{\mu}_i(\alpha_k^0,\theta^0),\theta^0\right)\left[\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0\right]$ and $\vec{\mu}_i(\alpha_k^0,\theta^0)$ lies between $\hat{\mu}_i(\alpha_k^0,\theta^0)$ and μ_i^0. Then using Assumption A1 and Lemma A.1, we can show that uniformly in i,

$$\begin{aligned}\hat{\mu}_i(\alpha_k^0,\theta^0)-\mu_i^0 &= -\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2T}\sum_{t=1}^{T}\hat{s}_{it,V}\right\}\\ &= -\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2T}\sum_{t=1}^{T}s_{it,V}\right\}+O_P(T^{-3/2}(\ln T)^9),\end{aligned}$$

where $\left[s_{it,V}\right]_j = (\frac{1}{T}\sum_{t=1}^{T}V_{it})^{\top}S_{iV}^{-1}V_{it,j}^{\mu\mu}S_{iV}^{-1}(\frac{1}{T}\sum_{t=1}^{T}V_{it})$. With this expression, we can readily show that

$$\begin{aligned}\mathbb{S}_{kNT,2} &= -\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^{T}U_{it}^{\mu}\left(\frac{1}{T}\sum_{t=1}^{T}V_{it}^{\mu}\right)^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}V_{it}+\frac{1}{2T}\sum_{t=1}^{T}s_{it,V}\right\}+o_P(1)\\ &= -\mathbb{S}_{kNT,2}(1)-\mathbb{S}_{kNT,2}(2)+o_P(1),\end{aligned}$$

where

$$\begin{aligned}
\mathbb{S}_{kNT,2}(1) &= \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} U_{it}^{\mu} \left(\frac{1}{T} \sum_{t=1}^{T} V_{it}^{\mu} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} V_{it} \\
&= \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} V_{it} + \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} \left[U_{it}^{\mu} - \mathbb{E}\left(U_{it}^{\mu}\right) \right] S_{iV}^{-1} \frac{1}{T} \sum_{t=1}^{T} V_{it} \\
&\quad + \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} S_{iU} \left[\left(\frac{1}{T} \sum_{t=1}^{T} V_{it}^{\mu} \right)^{-1} - S_{iV}^{-1} \right] \sum_{t=1}^{T} V_{it} + o_P(1) \\
&= \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} V_{it} + \frac{1}{\sqrt{N_k T^3}} \sum_{i\in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \left[U_{it}^{\mu} - \mathbb{E}\left(U_{it}^{\mu}\right) \right] S_{iV}^{-1} V_{is} \\
&\quad - \frac{1}{\sqrt{N_k T^3}} \sum_{i\in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} \left(V_{it}^{\mu} - S_{iV} \right) S_{iV}^{-1} V_{is} + o_P(1) \\
&= \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} V_{it} + \frac{1}{\sqrt{N_k T^3}} \sum_{i\in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{U}_{it}^{\mu} S_{iV}^{-1} V_{is} + o_P(1)
\end{aligned}$$

and

$$\mathbb{S}_{kNT,2}(2) = \frac{1}{2\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} U_{it}^{\mu} \left(\frac{1}{T} \sum_{t=1}^{T} V_{it}^{\mu} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} s_{it,V} = \frac{1}{2\sqrt{N_k T}} \sum_{i\in G_k^0} S_{iU} S_{iV}^{-1} R_{iV} + o_P(1).$$

where $[R_{iV}]_j = (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})^{\top} S_{iV}^{-1} S_{iV2,j} S_{iV}^{-1} (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})$. It follows that

$$\begin{aligned}
\mathbb{S}_{kNT,2} &= -\frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} V_{it} - \frac{1}{\sqrt{N_k T^3}} \sum_{i\in G_k^0} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{U}_{it}^{\mu} S_{iV}^{-1} V_{is} \\
&\quad - \frac{1}{2\sqrt{N_k T}} \sum_{i\in G_k^0} S_{iU} S_{iV}^{-1} R_{iV} + o_P(1).
\end{aligned} \tag{B.13}$$

Similarly we have

$$\begin{aligned}
\mathbb{S}_{\theta NT,2} &= -\frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} S_{iW} S_{iV}^{-1} V_{it} - \frac{1}{\sqrt{NT^3}} \sum_{i=1}^{N} \sum_{s=1}^{T} \sum_{t=1}^{T} \mathbb{W}_{it}^{\mu} S_{iV}^{-1} V_{is} \\
&\quad - \frac{1}{2\sqrt{NT}} \sum_{i=1}^{N} S_{iW} S_{iV}^{-1} R_{iW} + o_P(1),
\end{aligned} \tag{B.14}$$

where $[R_{iW}]_j = (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})^{\top} S_{iV}^{-1} S_{iW2,j} S_{iV}^{-1} (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})$. Combining (B.9)–(B.14) yields

$$\mathbb{S}_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i\in G_k^0} \sum_{t=1}^{T} \mathbb{U}_{it} - \mathbb{B}_{kNT} + o_P(1) \quad \text{and} \quad \mathbb{S}_{\theta NT} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{W}_{it} - \mathbb{B}_{\theta NT} + o_P(1).$$

By the Cramér-Wold device and Lindeberg-Feller central limit theorem, we can readily show that $((\mathbb{S}_{1NT} + \mathbb{B}_{1NT})^{\top}, \ldots, (\mathbb{S}_{K^0 NT} + \mathbb{B}_{K^0 NT})^{\top}, (\mathbb{S}_{\theta NT} + \mathbb{B}_{\theta NT})^{\top})^{\top}$ is asymptotically normally distributed with mean zero and variance-covariance matrix Ω. It follows that $\Xi_{NT}^{*} + \mathbb{H}_{NT}^{-1} \mathbb{B}_{NT} \xrightarrow{D} N\left(0, \mathbb{H}^{-1} \Omega \mathbb{H}^{-1}\right)$. This completes the proof of the theorem. ■

Proof of Lemma 4.1: By the spectral decompositions of $\boldsymbol{A}$ and $\boldsymbol{D}_N$, we have $\boldsymbol{A} = \boldsymbol{u}\Lambda\boldsymbol{u}^\top$ and $\boldsymbol{D}_N = \boldsymbol{U}_N\Sigma_N\boldsymbol{U}_N^\top = \boldsymbol{U}_{1,N}\Sigma_{1,N}\boldsymbol{U}_{1,N}^\top$. It follows that for any nonsingular matrix $\boldsymbol{S}$, we have

$$\begin{aligned}\boldsymbol{D}_N &= N^{-1}\boldsymbol{Z}_N\boldsymbol{A}\boldsymbol{Z}_N^\top = N^{-1}\boldsymbol{Z}_N\boldsymbol{u}\Lambda\boldsymbol{u}^\top\boldsymbol{Z}_N^\top \\ &= N^{-1}\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S}\boldsymbol{S}^{-1}\Lambda(\boldsymbol{S}^{-1})^\top\boldsymbol{S}^\top\boldsymbol{u}^\top\boldsymbol{Z}_N^\top = \boldsymbol{U}_{1,N}\Sigma_{1,N}\boldsymbol{U}_{1,N}^\top.\end{aligned}$$

Our goal is to find a nonsingular matrix $\boldsymbol{S}$ such that $\boldsymbol{U}_{1,N} = N^{-1/2}\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S}$ and $\Sigma_{1,N} = \boldsymbol{S}^{-1}\Lambda(\boldsymbol{S}^{-1})^\top$, which requires that $N^{-1}(\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S})^\top(\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S}) = I_{K^*}$ and $\boldsymbol{S}^{-1}\Lambda(\boldsymbol{S}^{-1})^\top$ should be diagonal. If such a matrix $\boldsymbol{S}$ exists, we must have

$$I_{K^*} = \boldsymbol{U}_{1,N}^\top\boldsymbol{U}_{1,N} = N^{-1/2}\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S},$$

yielding that $\boldsymbol{S} = (N^{-1/2}\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\boldsymbol{u})^{-1}$ provided that $\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\boldsymbol{u}$ is nonsingular. By construction, when $\boldsymbol{S} = (N^{-1/2}\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\boldsymbol{u})^{-1}$, $N^{-1}(\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S})^\top(\boldsymbol{Z}_N\boldsymbol{u}\boldsymbol{S}) = I_{K^*}$. In addition, we have

$$\begin{aligned}\boldsymbol{S}^{-1}\Lambda(\boldsymbol{S}^{-1})^\top &= N^{-1}\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\left(\boldsymbol{u}\Lambda\boldsymbol{u}^\top\right)\boldsymbol{Z}_N^\top\boldsymbol{U}_{1,N} = \boldsymbol{U}_{1,N}^\top\left(N^{-1}\boldsymbol{Z}_N\boldsymbol{A}\boldsymbol{Z}_N^\top\right)\boldsymbol{U}_{1,N} \\ &= \boldsymbol{U}_{1,N}^\top\boldsymbol{D}_N\boldsymbol{U}_{1,N} = \Sigma_{1,N},\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{D}_N = \boldsymbol{U}_{1,N}\Sigma_{1,N}\boldsymbol{U}_{1,N}^\top$ and $\boldsymbol{U}_{1,N}^\top\boldsymbol{U}_{1,N} = I_{K^*}$. So $\boldsymbol{S}^{-1}\Lambda(\boldsymbol{S}^{-1})^\top$ is diagonal and given by $\Sigma_{1,N}$ when $\boldsymbol{S} = (N^{-1/2}\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\boldsymbol{u})^{-1}$. The nonsingularity of $\boldsymbol{U}_{1,N}^\top\boldsymbol{Z}_N\boldsymbol{u}$ follows from the fact that $\boldsymbol{u}^\top\boldsymbol{u} = I_{K^*}$, $\boldsymbol{U}_{1,N}^\top\boldsymbol{U}_{1,N} = I_{K^*}$, and that the membership matrix $\boldsymbol{Z}_N$ is of full rank. This shows part (i)–(iii) of the lemma.

To prove (iv), we first show that the rows of $\boldsymbol{u}$ are distinct from each other. Suppose $\boldsymbol{u}$ has two identical rows, which are denoted as row k and row k'. We consider rows k, k' and columns k, k' of $\boldsymbol{A} \equiv \boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\top} = \boldsymbol{u}\Lambda\boldsymbol{u}^\top$:

$$\begin{bmatrix}c_1J_2 & \cdots & c_{K^*}J_2\end{bmatrix}\begin{bmatrix}\lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{K^*}\end{bmatrix}\begin{bmatrix}c_1J_2^\top \\ \vdots \\ c_{K^*}J_2^\top\end{bmatrix} = \left(\sum_{k=1}^{K^*}\lambda_k c_k^2\right)J_2J_2^\top,$$

where $J_2 = (1,1)^\top$ and c_k's are arbitrary scalars as long as $\boldsymbol{u}^\top\boldsymbol{u} = I_{K^*}$ is ensured. The last display has identical elements, which implies $\alpha_k^{0\top}\alpha_k^0 = \alpha_k^{0\top}\alpha_{k'}^0 = \alpha_{k'}^{0\top}\alpha_k^0 = \alpha_{k'}^{0\top}\alpha_{k'}^0$. This further implies that

$$\|\alpha_k^0 - \alpha_{k'}^0\|^2 = (\alpha_k^0 - \alpha_{k'}^0)^\top\left(\alpha_k^0 - \alpha_{k'}^0\right) = 0,$$

i.e., $\alpha_k^0 = \alpha_{k'}^0$ for $k \neq k'$, violating Assumption A2(i). Hence, we can conclude that the rows of $\boldsymbol{u}$ are distinct from each other. Since $\boldsymbol{S}$ is nonsingular, this further ensures that $\boldsymbol{u}\boldsymbol{S}$ has rows that are distinct from each other. Note that if z_i contains 1 in its kth element, then $z_i^\top\boldsymbol{u}\boldsymbol{S}$ is given by the kth row of $\boldsymbol{u}\boldsymbol{S}$. As a result, $z_i^\top\boldsymbol{u}\boldsymbol{S} = z_j^\top\boldsymbol{u}\boldsymbol{S}$ if and only $z_i = z_j$ for $i, j = 1, 2, \ldots, N$ because $\boldsymbol{u}\boldsymbol{S}$ has distinct rows. ■

Proof of Theorem 4.2: (i) We first prove that $\mathcal{K}_N = K^*$ w.p.a.1. Noting that $\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N = N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top - N^{-1}\boldsymbol{\beta}^0\boldsymbol{\beta}^{0\top} = N^{-1}[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)\boldsymbol{\beta}^{0\top} + \boldsymbol{\beta}^0(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0)^\top]$, we can readily show that

$$\|\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N\|^2 = O_P(N^{-1}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|^2) = O_P\left(T^{-1}\right). \tag{B.15}$$

By the perturbation theory for eigenvalue problems (e.g., Stewart and Sun (1990, p. 203), we have

$$\max_{1\leq\ell\leq N}|\tilde{\mu}_{\ell,N} - \mu_{\ell,N}| \leq \|\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N\| = O_P(T^{-1/2}),$$

where $\tilde{\mu}_{\ell,N}$ and $\mu_{\ell,N}$ denote the ℓth largest eigenvalues of $\tilde{\boldsymbol{D}}_N$ and $\boldsymbol{D}_N$, respectively. Since $\boldsymbol{D}_N$ has rank K^*, $\mu_{\ell,N} = 0$ and $\tilde{\mu}_{\ell,N} = O_P(T^{-1/2})$ for $\ell \geq K^* + 1$. This implies that $P\left(\mathcal{K}_N > K^*\right) \to 0$. By Assumptions

A5 and A2 and (4.3), $\tilde{\mu}_{\mathcal{K}_N} > \mu_{K^*N}/2 \geq c \min_{1\leq k\leq K^0} \tau_k/4 > 0$ w.p.a.1, implying that $P(\mathcal{K}_N < K^*) \to 0$. Consequently, we have $P(\mathcal{K}_N = K^*) \to 1$ as $(N,T) \to 1$.

(ii) We now prove the second part of the theorem. We find it is easy to consider the following singular value decompositions (SVDs) of $N^{-1/2}\boldsymbol{\beta}^0$ and $N^{-1/2}\tilde{\boldsymbol{\beta}}$:

$$N^{-1/2}\boldsymbol{\beta}^0 = \boldsymbol{U}_N \Sigma_N^{1/2} V_N^\top \text{ and } N^{-1/2}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{U}}_N \tilde{\Sigma}_N^{1/2} \tilde{V}_N^\top,$$

where $\boldsymbol{U}_N, \Sigma_N$ $\tilde{\boldsymbol{U}}_N$ and $\tilde{\Sigma}_N$ are as defined in Section 4.1, V_N is a $p \times p$ matrix such that $V_N^\top V_N = I_p$, $\tilde{V}_N$ is a $p \times p$ matrix such that $\tilde{V}_N^\top \tilde{V}_N = I_p$. Note that the (K^*+1)th, ..., pth diagonal elements of Σ_N are all zero, and the (K^*+1)th, ..., pth diagonal elements of $\tilde{\Sigma}_N$ are all $O_P\left(T^{-1/2}\right)$. We can decompose $\tilde{\boldsymbol{U}}_N$, $\tilde{\Sigma}_N$, and $\tilde{V}_N$ as follows: $\tilde{\boldsymbol{U}}_N = (\tilde{\boldsymbol{U}}_{1,N}, \tilde{\boldsymbol{U}}_{2,N})$, $\tilde{\Sigma}_N = \text{diag}(\tilde{\Sigma}_{1,N}, \tilde{\Sigma}_{2,N})$, and $\tilde{V}_N = (\tilde{V}_{1,N}, \tilde{V}_{2,N})$, where $\tilde{\boldsymbol{U}}_{1,N}$ is an $N \times K^*$ matrix, $\tilde{\Sigma}_{1,N}$ is a $K^* \times K^*$ diagonal matrix, and $\tilde{V}_{1,N}$ is a $p \times K^*$ matrix. Analogously, write $\boldsymbol{U}_N = (\boldsymbol{U}_{1,N}, \boldsymbol{U}_{2,N})$, $\Sigma = \text{diag}(\Sigma_{1,N}, \Sigma_{2,N})$, and $V_N = (V_{1,N}, V_{2,N})$, where $\Sigma_{2,N}$ is a matrix of zeros. Then

$$N^{-1/2}\boldsymbol{\beta}^0 = \boldsymbol{U}_N \Sigma_N^{1/2} V_N^\top = \boldsymbol{U}_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^\top \tag{B.16}$$

and

$$N^{-1/2}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{U}}_N \tilde{\Sigma}_N^{1/2} \tilde{V}_N^\top = \tilde{\boldsymbol{U}}_{1,N} \tilde{\Sigma}_{1,N}^{1/2} \tilde{V}_{1,N}^\top + \tilde{\boldsymbol{U}}_{2,N} \tilde{\Sigma}_{2,N}^{1/2} \tilde{V}_{2,N}^\top, \tag{B.17}$$

where $\tilde{\Sigma}_{1,N}^{1/2} - \Sigma_{1,N}^{1/2} = O_P(T^{-1/2})$, and $\tilde{\Sigma}_{2,N}^{1/2} = O_P(T^{-1/2})$. We consider the SVDs of $\tilde{\boldsymbol{U}}_{1,N}^\top \boldsymbol{U}_{1,N}$ and $\tilde{V}_{1,N}^\top V_{1,N}$:

$$\tilde{\boldsymbol{U}}_{1,N}^\top \boldsymbol{U}_{1,N} = A_{1N} \Theta_{1N} A_{2N}^\top \text{ and } \tilde{V}_{1,N}^\top V_{1,N} = B_{1N} \Theta_{2N} B_{2N}^\top,$$

where A_{1N}, A_{2N}, B_{1N} and B_{2N} are all orthogonal matrices, $\Theta_{1N} = \text{diag}\left(\cos\theta_{1,1}, \ldots, \cos\theta_{1,K^*}\right)$, $\Theta_{2N} = \text{diag}(\cos\theta_{2,1}, \ldots, \cos\theta_{2,K^*})$, $\theta_{1,1}, \ldots,$ and θ_{1,K^*} are the principal angles between the column spaces of $\tilde{\boldsymbol{U}}_{1,N}$ and $\boldsymbol{U}_{1,N}$, and $\theta_{2,1}, \ldots,$ and θ_{2,K^*} are the principal angles between the column spaces of $\tilde{V}_{1,N}$ and $V_{1,N}$. Let $O_N = A_{2N} A_{1N}^\top$ and $\tilde{O} = B_{2N} B_{1N}^\top$. Note that both O_N and $\tilde{O}$ are orthogonal matrices. Then by Theorem 4 in Yu, Wang, and Samworth (2015),

$$\|\tilde{\boldsymbol{U}}_{1,N} - \boldsymbol{U}_{1,N} O_N\| = O_P(N^{-1/2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|) = O_P(T^{-1/2}), \tag{B.18}$$

and

$$\|\tilde{V}_{1,N} - V_{1,N}\tilde{O}\| = O_P(N^{-1/2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\|) = O_P(T^{-1/2}). \tag{B.19}$$

To proceed, we first establish the connection between O_N and $\tilde{O}$ through $\Sigma_{1,N}^{1/2}$. Noting that $\|\tilde{\boldsymbol{U}}_{2,N} \tilde{\Sigma}_{2,N}^{1/2} \tilde{V}_{2,N}^\top\|$ $= O_P(\|\tilde{\Sigma}_{2,N}^{1/2}\|) = O_P(T^{-1/2})$, and by Theorem 3.1, the triangle inequality, the fact that O_N and $\tilde{O}$ are orthogonal matrices, equations (B.16)–(B.19), and the fact that $\boldsymbol{U}_{1,N}^\top \boldsymbol{U}_{1,N} = I_{K^*}$ and that $V_{1,N}^\top V_{1,N} = I_{K^*}$, we have

$$\begin{aligned}
O_P(T^{-1/2}) &= N^{-1/2}\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = \|\tilde{\boldsymbol{U}}_{1,N}\tilde{\Sigma}_{1,N}^{1/2}\tilde{V}_{1,N}^\top - \boldsymbol{U}_{1,N}\Sigma_{1,N}^{1/2}V_{1,N}^\top + \tilde{\boldsymbol{U}}_{2,N}\tilde{\Sigma}_{2,N}^{1/2}\tilde{V}_{2,N}^\top\| \\
&\geq \|\tilde{\boldsymbol{U}}_{1,N}\Sigma_{1,N}^{1/2}\tilde{V}_{1,N}^\top - \boldsymbol{U}_{1,N}\Sigma_{1,N}^{1/2}V_{1,N}^\top\| - O_P(T^{-1/2}) \\
&= \|\tilde{\boldsymbol{U}}_{1,N}O_N^\top O_N \Sigma_{1,N}^{1/2}\tilde{O}^\top\tilde{O}\tilde{V}_{1,N}^\top - \boldsymbol{U}_{1,N}\Sigma_{1,N}^{1/2}V_{1,N}^\top\| + O_P(T^{-1/2}) \\
&= \|\boldsymbol{U}_{1,N}O_N\Sigma_{1,N}^{1/2}\tilde{O}^\top V_{1,N}^\top - \boldsymbol{U}_{1,N}\Sigma_{1,N}^{1/2}\tilde{O}\tilde{O}^\top V_{1,N}^\top\| + O_P(T^{-1/2}) \\
&= \|\boldsymbol{U}_{1,N}(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O})\tilde{O}^\top V_{1,N}^\top\| + O_P(T^{-1/2}) \\
&= \left[\text{tr}\left(\boldsymbol{U}_{1,N}(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O})\tilde{O}^\top V_{1,N}^\top V_{1,N}\tilde{O}(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O})^\top \boldsymbol{U}_{1,N}^\top\right)\right]^{1/2} + O_P(T^{-1/2}) \\
&= \left[\text{tr}\left((O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O})(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O})^\top\right)\right]^{1/2} + O_P(T^{-1/2}) \\
&= \|O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O}\| + O_P(T^{-1/2}).
\end{aligned}$$

It follows that $\|O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O}\| = O_P(T^{-1/2})$.

Recall that $\tilde{u}_i^\top = \left(\tilde{u}_{1,i}^\top, \tilde{u}_{2,i}^\top\right)$ and $u_i^\top = \left(u_{1,i}^\top, u_{2,i}^\top\right)$ denote the ith row of $\tilde{\boldsymbol{U}}_N = (\tilde{\boldsymbol{U}}_{1,N}, \tilde{\boldsymbol{U}}_{2,N})$ and $\boldsymbol{U}_N = (\boldsymbol{U}_{1,N}, \boldsymbol{U}_{2,N})$, respectively. Notice that uniformly in i

$$O_P(N^{-1/2}) = N^{-1/2}\tilde{\beta}_i^\top = \tilde{u}_{1,i}^\top\tilde{\Sigma}_{1,N}^{1/2}\tilde{V}_{1,N}^\top + \tilde{u}_{2,i}^\top\tilde{\Sigma}_{2,N}^{1/2}\tilde{V}_{2,N}^\top.$$

We post-multiply both sides of the above expression by $\tilde{V}_{1,N}$ and apply the fact that $\tilde{V}_{1,N}^\top\tilde{V}_{1,N} = I_{K^*}$ and that $\tilde{V}_{2,N}^\top\tilde{V}_{1,N} = 0$ to obtain

$$O_P(N^{-1/2}) = N^{-1/2}\tilde{\beta}_i^\top\tilde{V}_{1,N} = \tilde{u}_{1,i}^\top\tilde{\Sigma}_{1,N}^{1/2}.$$

It follows that uniformly in i we have $\tilde{u}_{1,i} = \tilde{\Sigma}_{1,N}^{-1/2}O_P(N^{-1/2}) = O_P(N^{-1/2})$. That is, $\max_{1\le i\le N}\|\tilde{u}_{1,i}\| = O_P(N^{-1/2})$. Next, we compare $\tilde{u}_{1,i}$ and $u_{1,i}$ through $\tilde{\beta}_i$ and β_i^0. By Theorem 3.1 and the fact that $\tilde{\Sigma}_{1,N}^{1/2} - \Sigma_{1,N}^{1/2} = O_P\left(T^{-1/2}\right)$ and that $\max_{1\le i\le N}\|\tilde{u}_{1,i}\| = O_P(N^{-1/2})$, we have uniformly in i,

$$\begin{aligned}
O_P(T^{-1/2}(\ln T)^3) &= (\tilde{\beta}_i - \beta_i^0)^\top\tilde{V}_{1,N} = \sqrt{N}\tilde{u}_{1,i}^\top\tilde{\Sigma}_{1,N}^{1/2} - \sqrt{N}u_{1,i}^\top\Sigma_{1,N}^{1/2}V_{1,N}^\top\tilde{V}_{1,N} \\
&= \sqrt{N}\tilde{u}_{1,i}^\top\Sigma_{1,N}^{1/2} - \sqrt{N}u_{1,i}^\top\Sigma_{1,N}^{1/2}V_{1,N}^\top\tilde{V}_{1,N} + \sqrt{N}\tilde{u}_{1,i}^\top(\tilde{\Sigma}_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}) \\
&= \sqrt{N}(\tilde{u}_{1,i} - O_N^\top u_{1,i})^\top\Sigma_{1,N}^{1/2} + \sqrt{N}u_{1,i}^\top(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}V_{1,N}^\top\tilde{V}_{1,N}) + O_P(T^{-1/2}).
\end{aligned}$$

It follows that uniformly in i,

$$\begin{aligned}
\sqrt{N}\left\|\tilde{u}_{1,i} - O_N^\top u_{1,i}\right\| &= \| - \Sigma_{1,N}^{-1/2}\sqrt{N}(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}V_{1,N}^\top\tilde{V}_{1,N})^\top u_{1,i} + O_P(T^{-1/2}(\ln T)^3)\| \\
&\le \|\Sigma_{1,N}^{-1/2}(O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O})^\top\sqrt{N}u_{1,i}\| \\
&\quad + \|\Sigma_{1,N}^{-1/2}[\Sigma_{1,N}^{1/2}V_{1,N}^\top(\tilde{V}_{1,N} - V_{1,N}\tilde{O})]^\top\sqrt{N}u_{1,i}\| + O_P(T^{-1/2}(\ln T)^3) \\
&\le \|\Sigma_{1,N}^{-1/2}\| \cdot \|O_N\Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2}\tilde{O}\| \cdot \|\sqrt{N}u_{1,i}\| + O_P(T^{-1/2}(\ln T)^3) \\
&= O_P(1)\, O_P(T^{-1/2}(\ln T)^3)O_P(1) + O_P(T^{-1/2}(\ln T)^3) = O_P(T^{-1/2}(\ln T)^3).
\end{aligned}$$

This completes the proof of the theorem. ■

Proof of Theorem 4.3: The proof is similar to that of Theorem 3.2. The major difference is that we now work in the eigenspace $\tilde{\boldsymbol{U}}_{1,N}$ instead of the preliminary estimate matrix $\tilde{\boldsymbol{\beta}}$, and now each row of $\sqrt{N}\tilde{\boldsymbol{U}}_{1,N}$ is consistent with the corresponding row of $\sqrt{N}\boldsymbol{U}_{1,N}O$, which contains the group membership information. Here O denotes the probability limit of O_N. Now $\sqrt{N}\tilde{u}_{1,i} = O_P(1)$ and$\sqrt{N}O^\top u_{1,i} = O(1)$ for each i, and they play the roles of $\tilde{\beta}_i$ and β_i^0 in the proof of Theorem 3.2, respectively. Furthermore, the result in Theorem 4.2(ii) implies the consistency of $\sqrt{N}\tilde{u}_{1,i}$ is uniform in all individuals, which is sufficient for us to identify all the individuals group membership. ■

Proof of Theorem 4.4: Given the result in Theorem 4.3, the proof of the theorem follows that of Theorem 3.3 and thus omitted. ■

Proof of Theorem 4.5: Given the consistency of $\tilde{\mathcal{G}}$ with $\mathcal{G}^0$ by Theorems 4.3–4.4, the proof of the theorem is completely analogous to that of Theorem 3.4 and thus omitted. ■

REFERENCES

Abrevaya, J., Shen, S., 2014. Estimation of censored panel data models with slope heterogeneity. Journal of Applied Econometrics 29, 523–548.

Alan, S., Honoré, B. E., Hu, L., Leth-Petersen, S., 2014. Estimation of panel data regression models with two-sided censoring or truncation. Journal of Econometric Methods 3, 1–20.

Alessie, R., Hochguertel, S., Van Soest, A., 2002. Household Portfolios in the Netherlands. MIT Press, Cambridge.

Ando, T., Bai, J., 2016. Panel data models with grouped factor structure under unknown group membership. Journal of Applied Econometrics 31, 163–191.

Bai, J., 1997. Estimating multiple breaks one at a time. Econometric Theory 13, 315–352.

Bhatia, R., 1997. Matrix Analysis. Springer-Verlag, New York.

Bester, C. A., Hansen, C. B., 2016. Grouped effects estimators in fixed effects models. Journal of Econometrics 190, 197–208.

Bonhomme, S., Manresa, E., 2015. Grouped patterns of heterogeneity in panel data. Econometrica 83, 1147–1184.

Breiman, L., Friedman, J., Stone, C. J., Olshen, R. A., 1984. Classification and Regression Trees. CRC Press.

Browning, M., Carro, J., 2007. Heterogeneity and microeconometrics modeling. Econometric Society Monographs 43.

Cocco, J. F., Gomes, F. J., Maenhout, P. J., 2005. Consumption and portfolio choice over the life cycle. Review of Financial Studies 18, 491–533.

Curcuru, S., Heaton, J., Lucas, D., Moore, D., 2004. Heterogeneity and portfolio choice: theory and evidence. Handbook of Financial Econometrics 1, 337–382.

Chen, M., 2016. Estimation of nonlinear panel models with multiple unobserved effects. Working Paper, Department of Economics, University of Warwick.

Chen, M., Fernandez-Val, I., Weidner, M., 2014. Nonlinear panel models with interactive effects. Working Paper, Department of Economics, University of Warwick.

Davis, C., Kahan, W. M., 1970. The rotation of eigenvectors by a perturbation: III. SIAM Journal on Numerical Analysis 7, 1–46.

Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effect models. Review of Economic Studies 82, 991–1030.

Fan, J., Lv, J., Qi, L., 2011. Sparse high dimensional models in economics. Annual Review of Economics 3, 291–317.

Hahn, J., Kuersteiner, G., 2011. Bias reduction for dynamic nonlinear panel models with fixed effects. Econometric Theory 27, 1152–1191.

Hahn, J., Newey, W., 2004. Jackknife and analytical bias reduction for nonlinear panel models. Econometrica 72, 1295–1319.

Hall, P., Heyde, C. C., 1980. Martingale Limit Theory and Its Application. Academic Press.

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

Hsiao, C., Tahmiscioglu, A. K., 1997. A panel analysis of liquidity constraints and firm investment. Journal of the American Statistical Association 92, 455–465.

Hu, L., 2002. Estimation of a censored dynamic panel data model. Econometrica 70, 2499–2517.

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

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

Lin, C.-C., Ng, S., 2012. Estimation of panel data models with parameter heterogeneity when group membership is unknown. Journal of Econometric Methods 1, 42–55.

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

Okui, R., Wang, W., 2017. Heterogeneous structural breaks in panel data models. Working Paper, Erasmus School of Economics, Erasmus Universiteit Rotterdam.

Pesaran, M. H., Shin, Y., Smith, R. P., 1999. Pooled mean group estimation of dynamic heterogeneous panels. Journal of the American Statistical Association 94, 621–634.

Phillips, P. C. B., Sul, D., 2007. Transition modeling and econometric convergence tests. Econometrica 75, 1771–1855.

Polkovnichenko, V., 2007. Life-cycle portfolio choice with additive habit formation preferences and uninsurable labor income risk. Review of Financial Studies 20, 83–124.

Rohe, K., Chatterjee, S., Yu, B., 2011. Spectral clustering and the high-dimensional stochastic blockmodel. Annals of Statistics 39, 1878–1915.

Samuelson, P. A., 1969. Lifetime portfolio selection by dynamic stochastic programming. Review of Economics and Statistics, 239–246.

Sarafidis, V., Weber, N., 2015. A partially heterogeneous framework for analyzing panel data. Oxford Bulletin of Economics and Statistics 77, 274–296.

Shen, X., Huang, H.-C., 2010. Grouping pursuit through a regularization solution surface. Journal of the American Statistical Association 105, 727–739.

Stewart, G. W., Sun, J. G., 1990. Matrix Perturbation Theory. Academic Press.

Su, L., Chen, Q., 2013. Testing homogeneity in panel data models with interactive fixed effects. Econometric Theory 29, 1079–1135.

Su, L., Ju, G., 2017. Identifying latent grouped patterns in panel data models with interactive fixed effects. Journal of Econometrics, forthcoming.

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

Su, L., Shi, Z., Phillips, P. C. B., 2016b. Supplement to "Identifying latent structures in panel data", Econometrica Supplemental Material 84, http://dx.doi.org/10.3982/ECTA12560.

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

Subramanian, A., Wei, S.-J., 2007. The WTO promotes trade, strongly but unevenly. Journal of International Economics 72, 151–175.

von Luxburg, U., 2007. A tutorial on spectral clustering. Statistics and Computing 17, 395–416.

Wang, W., Phillips, P. C. B., Su, L., 2017. Homogeneity pursuit in panel data models: theory and applications. Working Paper, Singapore Management University.

Yu, Y., Wang, T., Samworth, R. J., 2015. A useful variant of the Davis-Kahan theorem for statisticians. Biometrika 102, 315–323.