2-2021

Identifying latent group structures in nonlinear panels

Wuyi WANG
Jinan University - China

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

Identifying Latent Group Structures in Nonlinear Panels[*]

Wuyi Wang[a] and Liangjun Su[b]

a Institute for Economic and Social Research, Jinan University [b] School of Economics, Singapore Management University

Published in Journal of Econometrics, 2021 February, 220 (2), 272-295

DOI: 10.1016/j.jeconom.2020.04.003

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

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.

[*] The authors thank the guest editors and two anonymous referees for many constructive comments on the paper. They also thank the conference participants at the 23rd International Panel Data Conference (2017, Thessaloniki), the 13th Symposium on Econometric Theory and Application (SETA2017, 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. Wang acknowledges support from the 111 Project of China (Project Number: B18026). 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.

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) and 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 et al., 2015). In the previous two examples, the group structures are latent. In some other studies, the group structures are assumed to be observable. 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 et al., 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.

The panel structure model is also closely related to the subgroup analysis in the statistics literature. Statisticians are interested in identifying the latent subgroups in order to design group-specific treatments in clinical trials, marketing strategies and so on. Shen and He (2015) use a structured logistic-normal mixture model to test the existence of subgroups and obtain predictive scores for the subgroup membership at the same time. Ma and Huang (2017) propose a penalized approach for subgroup analysis by applying concave penalty functions to pairwise differences of the intercepts in a regression model. Radchenko and Mukherjee (2017) study the asymptotic properties of a convex clustering method, which is adapted from the K-means algorithm.

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) and Ando and Bai (2016). The second popular approach is based on the classifier-Lasso (C-Lasso) that has been recently proposed by Su et al. (2016a, SSP hereafter) and extended in Su and Ju (2018), Su et al. (2019), and Wang et al. (2019). 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 et al. (2018, WPS hereafter) extend the CARDS algorithm of Ke et al. (2015) to the panel data framework to identify the group structure of slope parameters.

Recently, Ke et al. (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 et al., 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 contain 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 WPS, both papers rely upon some preliminary consistent estimates of the regression coefficients to obtain ordered statistics along each component and then proceed to estimate the latent panel structure. The main differences lie in two aspects. First, WPS rely on the ordered segmentations to construct the Lasso-type penalties for individuals within the same segment and for those in neighboring segments, while we employ the ordering simply for the purpose of extending the SBSA from the time series structural change literature to identifying the latent group structure in our setup. Second, WPS need to specify at least three tuning parameters, one for the between-segment penalty, one for the within-segment penalty, and the other one for controlling the number of the segments. In sharp contrast, we do not need to specify any tuning parameter once the number of groups is given.

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 needs 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 et al. (1999)) while others' are not. Besides, when a regressor (e.g., employment status) is time-invariant for some individuals, we have no choice but to assume its slope coefficient is homogeneous across individuals in the linear-type panel data models.[1]

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 et al. (2011)). Using a useful variant of the deep Davis-Kahan $\sin\theta$ theorem *a la* Yu et al. (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 responds 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 on 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. To save space, all proofs are relegated to the online supplementary appendix.

Notation. $\mathbb{R}^n$ ($\mathbb{N}^n$) denotes the n-dimensional Euclidean (natural number) space. 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 $\overset{D}{\to}$ and $\overset{P}{\to}$ denote convergence in distribution and in probability, respectively.

[1] In the empirical application studied in this paper, some people change from the work status to the retirement status while others stay in the work or retirement status over the whole observed time periods. In other words, the retirement dummy changes from 0 to 1 for some individuals and remains fixed to be either 0 or 1 over the period of study for the other individuals.

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 that may contain the lagged dependent variables (e.g., $y_{i,t-1}$), ε_{it} is the idiosyncratic shock, μ_i is an $r \times 1$ vector of nuisance parameters (e.g., the fixed effects), θ is a $q \times 1$ vector of parameters that are 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 $|G|$ denotes the cardinality of the set G. 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

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

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

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

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 $\beta_{i,j}^0$, $\alpha_{k,j}^0$ and $\tilde{\beta}_{i,j}$ denote the jth element of β_i^0, α_k^0 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 $\beta_{i,j}^0$'s are not identical across i for some j, then finding the homogeneity among $\beta_{i,j}^0$'s is equivalent to finding the "break points" among the ordered version of $\beta_{i,j}^0$'s. When $\tilde{\beta}_{i,j}$'s are consistent estimates of $\beta_{i,j}^0$'s, we expect the "break points" in the ordered $\beta_{i,j}^0$'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 $\beta_{i,j}^0$'s. For example, suppose $K^0 = 3$, $\alpha_{1,j}^0 < \alpha_{2,j}^0 < \alpha_{3,j}^0$, and N_1 (resp. N_2 and $N - N_1 - N_2$) $\beta_{i,j}^0$'s take value $\alpha_{1,j}^0$ (resp. $\alpha_{2,j}^0$ and $\alpha_{3,j}^0$). 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 $\alpha_{1,j}^0$, $\alpha_{2,j}^0$, and $\alpha_{3,j}^0$, 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 $\alpha_{1,j}^0$, $\alpha_{2,j}^0$, and $\alpha_{3,j}^0$ 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

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

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

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

which measures the variation in $\mathcal{S}_{i,l}(j)$ in the presence of a conjectured break point at m. Since K^0 is typically unknown, we have to pick up a large enough number $K^{\max}$ such that $1 \le K^0 \le K^{\max}$. 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)[2]

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 $\hat{j}_1 = \arg\max_{1\le j\le p}\hat{V}_{1,N}(j)$. Given $\hat{j}_1$, we solve the following minimization problem

   $$\hat{m}_1 \equiv \underset{1\le m<N}{\arg\min}\,\hat{S}_{1,N}(\hat{j}_1, m),$$

   which is an estimated break point. Now we have two segments – $\hat{G}_1(2) = \mathcal{S}_{1,\hat{m}_1}(\hat{j}_1)$ and $\hat{G}_2(2) = \mathcal{S}_{\hat{m}_1+1,N}(\hat{j}_1)$.
3. When $K \ge 3$, 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

   $$\hat{j}_{K-1} \equiv \underset{1\le j\le p}{\arg\max}\sum_{k=1}^{K-1}\hat{V}_{\hat{m}_{k-1}+1,\hat{m}_k}(j),$$
   $$\hat{m}_{K-1}(k) \equiv \underset{\hat{m}_{k-1}+1\le m<\hat{m}_k}{\arg\min}\hat{S}_{\hat{m}_{k-1}+1,\hat{m}_k}(\hat{j}_{K-1}, m) \text{ for } k = 1, \ldots, K-1,$$

   where $\hat{m}_0 = 0$, $\hat{m}_{K-1} = N$, and we suppress the dependence of $\hat{m}_{K-1}(k)$ on $\hat{j}_{K-1}$. 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 $k = 1, \ldots, K-1$,

   $$\hat{S}_{K-1}(k) \equiv \sum_{i\in\hat{G}_{k1}(K-1)}\left[\tilde{\beta}_{i,\hat{j}_{K-1}} - \bar{\beta}_{\hat{G}_{k1}(K-1)}\left(\hat{j}_{K-1}\right)\right]^2 + \sum_{i\in\hat{G}_{k2}(K-1)}\left[\tilde{\beta}_{i,\hat{j}_{K-1}} - \bar{\beta}_{\hat{G}_{k2}(K-1)}\left(\hat{j}_{K-1}\right)\right]^2$$
   $$+ \sum_{1\le l\le K-1, l\ne k}\sum_{i\in\hat{G}_l(K-1)}\left[\tilde{\beta}_{i,\hat{j}_{K-1}} - \bar{\beta}_{\hat{G}_l(K-1)}\left(\hat{j}_{K-1}\right)\right]^2,$$

   where, e.g., $\bar{\beta}_{\hat{G}_{k1}(K-1)}\left(\hat{j}_{K-1}\right) = |\hat{G}_{k1}(K-1)|^{-1}\sum_{i\in\hat{G}_{k1}(K-1)}\tilde{\beta}_{i,\hat{j}_{K-1}}$. Note that $\hat{S}_{K-1}(k)$ measures the subsegment variation of $\{\tilde{\beta}_{i,\hat{j}_{K-1}}\}_{i=1}^{N}$ when an additional potential break point $\hat{m}_{K-1}(k)$ is detected in the set $\hat{G}_k(K-1)$. Let

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

[2] 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.

which is the estimated segment number based on which we find the new break point. 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 $\hat{\mathcal{G}}\left(K^0\right) \equiv \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{K^0}\}$ 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 $\hat{\boldsymbol{\beta}}(K) = (\hat{\beta}_1(K), \ldots, \hat{\beta}_N(K))^\top$, $\hat{\boldsymbol{\mu}}(K) = (\hat{\mu}_1(K), \ldots, \hat{\mu}_N(K))^\top$, $\hat{\alpha}(K) = \left(\hat{\alpha}_1(K), \ldots, \hat{\alpha}_K(K)\right)^\top$, 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 function of (N, T). It plays the role of $\ln(NT)/(NT)$ in the use of BIC in the panel setup. Let

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

be the estimated number of groups and the estimated group structure, respectively. We will show that

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

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 estimators 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

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

where $W_{it}^{\theta}(\theta) = W_i^{\theta}(w_{it}; \gamma_i(\theta), \theta)$ and $W_{it}^{\gamma_i}(\theta) = W_i^{\gamma_i}(w_{it}; \gamma_i(\theta), \theta)$. 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 $\inf_{\theta:\|\theta-\theta^0\|>\eta} \frac{1}{N}\sum_{i=1}^{N} \left[\Psi_i(\gamma_i(\theta), \theta) - \Psi_i(\gamma_i(\theta^0), \theta^0)\right] > \epsilon$, where $\Psi_i(\gamma_i, \theta) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\varphi(w_{it}; \gamma_i, \theta)]$.

(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 $\varsigma_i, \varsigma_i' \in \Upsilon$ 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) $NT^{1-\kappa/2} \to c \in [0, \infty)$, $(p/T)(\ln T)^6 \to 0$ and $q = O(p)$ as $(N, T) \to \infty$.

Assumption A1(i)–(v) parallel Assumption 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. Note that Assumption A1(i) allows the dynamic panel data model. 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) imposes conditions on N, T, p, and q. It requires that N, p, and q should 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 condition $q = O(p)$ is imposed to facilitate the presentation below because in this case the estimation of θ^0 will not affect the L_2-convergence rate of $\tilde{\gamma}_i$'s.

The following theorem establishes the consistency of the preliminary estimates $\tilde{\theta}$ and $\tilde{\gamma}_i$.

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

The proof of Theorem 3.1 is rather complicated and relegated to the appendix. The results are applied to both static and dynamic panel data models. The rate in Theorem 3.1(iii) is not optimal. In fact, following Su et al. (2016b), SSPb hereafter) we can establish that $P(\max_{1\le i\le N}\left\|\tilde{\gamma}_i - \gamma_i^0\right\| \ge C(p/T)^{1/2}(\ln T)^3) = o(N^{-1})$ 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 $o(N^{-1})$ or relax the moment conditions in Assumption A1(iv).

Note that SSP also obtains preliminary rates of convergence for their Lasso estimators. Our results in Theorem 3.1(ii) and (iv) parallel those in Theorem 2.1(i)–(ii) of SSP. The major difference is that we allow for the presence of a common parameter (θ) and permit both p and q to diverge to infinity whereas SSP does not allow the presence of a common parameter in their basic model and only focus on the fixed p case. As a result, the convergence rates of our estimators $\tilde{\theta}$ and $\tilde{\gamma}_i$ in terms of Frobenius norm depend on q and p, respectively, while that of SSP's estimates of the regression coefficient are affected by the choice of a tuning parameter in their C-Lasso procedure.

3.2. Consistency of classification

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

We add two assumptions.

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

(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) $p^{3/2}N^{1/2}(\ln N)^9/T \to 0$ as $(N, T) \to \infty$.

(ii) As $(N, T) \to \infty$, $\min_{1 \le K < K^0} \min_{\mathcal{G}(K)} \hat{\sigma}^2_{\mathcal{G}(K)} \xrightarrow{P} \bar{\sigma}^2 > \sigma_0^2$, where $\sigma_0^2 \equiv \lim_{(N,T)\to\infty} 2(NT)^{-1} \sum_{k=1}^{K^0} \sum_{i \in G_k^0} \sum_{t=1}^{T} \mathbb{E}\varphi(w_{it}; \alpha_k^0, \mu_i^0, \theta^0)$.

(iii) $p\rho_{NT} \to 0$ 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 $\min_{1 \le k < k' \le K^0} \|\alpha_k^0 - \alpha_{k'}^0\|$ to shrink to zero at a rate slower than $(p/T)^{1/2} (\ln T)^3$. 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 a nonnegligible ratio of members asymptotically. Assumption A3(i) parallels Assumption A2(ii) in SSP and 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 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_0^2.

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

Theorem 3.2 (*Classification Consistency*)**.** *Suppose that 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 $P(\hat{\mathcal{G}}\left(K^0\right) = \mathcal{G}^0) \to 1$ 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 that 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*). $p\rho_{NT} \to 0$ 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).** $p\rho_{NT} \to 0$ 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 denote 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 $U_{it,j}^{\mu}\equiv\partial U_j(w_{it};\beta_i^0,\mu_i^0,\theta^0)/\partial\mu_i^{\top}$, $U_{it,j}^{\mu\mu}\equiv\partial^2U_j(w_{it};\beta_i^0,\mu_i^0,\theta^0)/\partial\mu_i\partial\mu_i^{\top}$, and similarly for $W_{it,j}^{\mu}$, $V_{it,j}^{\mu\mu}$ and $W_{it,j}^{\mu\mu}$. Define $S_{iU}\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(U_{it}^{\mu})$, $S_{iV}\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(V_{it}^{\mu})$, $S_{iW}\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(W_{it}^{\mu})$, $S_{iU2,j}\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(U_{it,j}^{\mu\mu})$, $S_{iV2,j}\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(V_{it,j}^{\mu\mu})$, $S_{iW2,j}\equiv\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(W_{it,j}^{\mu\mu})$, $\mathbb{U}_{it}\equiv U_{it}-S_{iU}S_{iV}^{-1}V_{it}$, $\mathbb{U}_{it}^{\mu}\equiv U_{it}^{\mu}-S_{iU}S_{iV}^{-1}V_{it}^{\mu}$, $\mathbb{W}_{it}\equiv W_{it}-S_{iW}S_{iV}^{-1}V_{it}$, $\mathbb{W}_{it}^{\mu}\equiv W_{it}^{\mu}-S_{iW}S_{iV}^{-1}V_{it}^{\mu}$, $\Omega_{iT,\beta\beta}\equiv\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}(\mathbb{U}_{is}\mathbb{U}_{it}^{\top})$, $\Omega_{iT,\beta\theta}\equiv\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}(\mathbb{U}_{is}\mathbb{W}_{it}^{\top})$, and $\Omega_{iT,\theta\theta}\equiv\frac{1}{T}\sum_{s=1}^{T}\sum_{t=1}^{T}\mathbb{E}(\mathbb{W}_{is}\mathbb{W}_{it}^{\top})$. Next, let $\mathbb{B}_{NT}\equiv(\mathbb{B}_{1NT}^{\top},\ldots,\mathbb{B}_{K^0NT}^{\top},\mathbb{B}_{\theta NT}^{\top})^{\top}=((\mathbb{B}_{1,1NT}-\mathbb{B}_{2,1NT})^{\top},\ldots,(\mathbb{B}_{1,K^0NT}-\mathbb{B}_{2,K^0NT})^{\top},(\mathbb{B}_{1,\theta NT}-\mathbb{B}_{2,\theta NT})^{\top})^{\top}$, where

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

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

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

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

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

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

where

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

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

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

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

Let $\mathbb{H}_{NT}\equiv\mathbb{H}_{NT}\left(\boldsymbol{\beta}^0,\theta^0\right)$. Note that $\mathbb{H}_{NT}^{-1}\mathbb{B}_{NT}$ and $\mathbb{H}_{NT}^{-1}\Omega_{NT}(\mathbb{H}_{NT}^{-1})^{\top}$ are associated with the asymptotic bias and variance of our estimators, respectively.

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) $\mathbb{H}\equiv\lim_{(N,T)\to\infty}\mathbb{E}[\mathbb{H}_{NT}]$ 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*

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

where $D_{NT}=\operatorname{diag}(\sqrt{N_1T}I_p,\ldots,\sqrt{N_{K^0}T}I_p,\sqrt{NT}I_q)$, $\mathbf{S}$ is an arbitrary $k\times\left(pK^0+q\right)$ selection matrix with full rank k, and $k\in[1,pK^0+q]$ is a fixed positive integer.

Note that we specify a selection matrix $\mathbf{S}$ in Theorem 3.4. It is not needed if both p and q remain fixed as $(N,T)\to\infty$. That is, one can replace $\mathbf{S}$ by I_{pK^0+q} in this case to obtain the joint normality of $\hat{\delta}_{NT}\equiv((\hat{\alpha}_1-\alpha_1^0)^{\top},\ldots,(\hat{\alpha}_{K^0}-\alpha_{K^0}^0)^{\top},(\hat{\theta}-\theta^0)^{\top})^{\top}$. When p or q or both pass to infinity as $(N,T)\to\infty$, the dimension of $\hat{\delta}_{NT}$ also diverges to infinity at the rate $p+q$ so that we cannot derive its asymptotic normality directly. We follow the literature on inferences with a diverging number of parameters (e.g., Lam and Fan, 2008; Qian and Su, 2016) and prove the asymptotic normality for any arbitrary linear combinations of elements of $\hat{\delta}_{NT}$.

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 expansions 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 formulas are a little bit more complicated than theirs due to the presence of the common parameter θ. In the absence of θ, both formulas 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. Following Hahn and Newey (2004) and Hahn and Kuersteiner (2011) and SSP, it is easy to show that elements of $\mathbb{B}_{NT}$ are $o_P(1)$ and the bias term can be removed in (3.3) if the model is a linear model and all regressors are strictly exogenous. In the more general case where the model is either nonlinear or contains lagged dependent variables, the elements of $\mathbb{B}_{NT}$ are $O_P(\sqrt{N/T})$.

To make the 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 formulas 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 individuals' 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 follow directly from the previous section.

4.1. Spectral decomposition

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

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

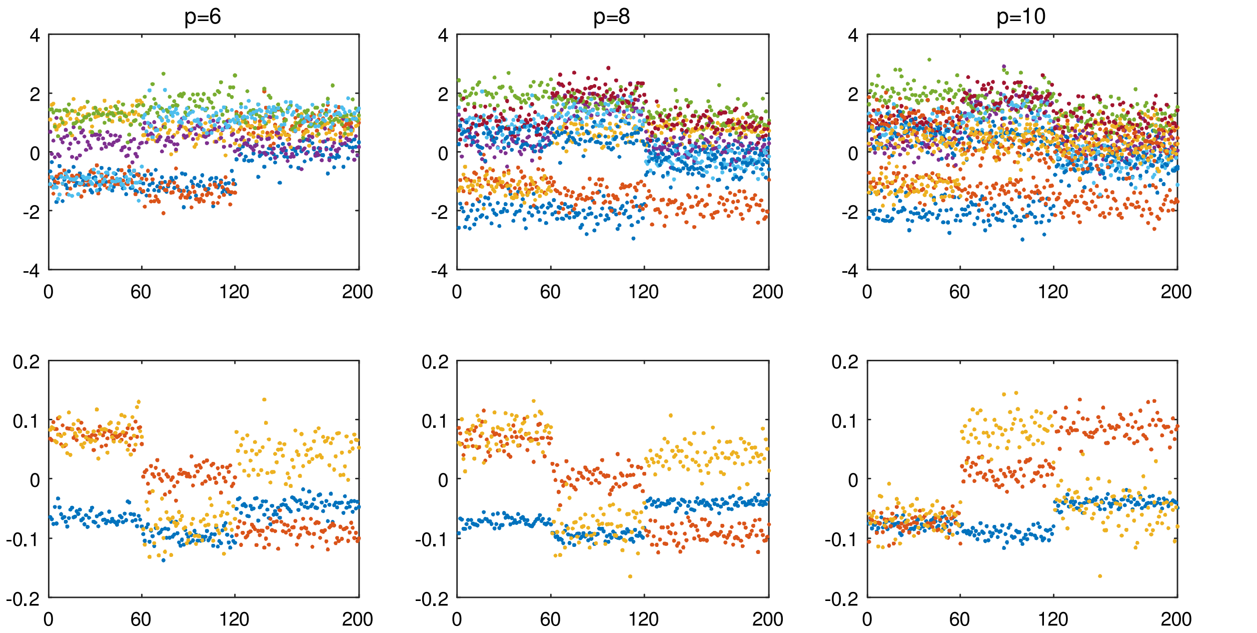

Fig. 1. Comparison of the plots of the p columns in the preliminary estimates $\tilde{\boldsymbol{\beta}}$ with the three eigenvectors of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^{\top}$ associated with its three largest eigenvalues when $p > K^0$: row 1 for preliminary estimates and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

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 an 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. Comparing with the SBM, $\boldsymbol{Z}_N$ stores the individuals' group membership in our model and nodes' community membership in an SBM. The matrix $\boldsymbol{A}$ here is analogous to the probability matrix that contains the probability of edges within and between blocks in an 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 an SBM. It has been suggested that the eigenvectors corresponding to the few largest eigenvalues of a certain matrix associated with the adjacency matrix reveal the clusters of interest. For example, Rohe et al. (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 two examples 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 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 Fig. 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 $p = 10$).

Example 4.2 (*Linear Dependence*)**.** In this example we consider a case that the rows of the $p \times K^0$ matrix $\boldsymbol{\alpha}^{0\top}$ are linearly dependent. Let $p = 2$ and $K^0 = 3$. The true group-specific parameters are

$$\boldsymbol{\alpha}^{0\top} = (\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1.4 \\ 1.4 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix} \right),$$

where $(\alpha_{11}^0, \alpha_{21}^0, \alpha_{31}^0)^\top$ and $(\alpha_{12}^0, \alpha_{22}^0, \alpha_{32}^0)^\top$ are linearly dependent. Now, the eigenvector associated with the largest eigenvalue of $N^{-1}\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\beta}}^\top$ contains all the useful individual group identity information in $\tilde{\boldsymbol{\beta}}$ and essentially serves as a "signal enhancement". As a result, such an eigenvector can reveal the true group information much more clearly than $\tilde{\boldsymbol{\beta}}$ itself.

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 $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_{K^*})$ 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. 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 $\Sigma_N = \mathrm{diag}(\mu_{1N}, \ldots, \mu_{K^*N}, 0, \ldots, 0)$ is a $p \times p$ matrix that contains the eigenvalues of $\boldsymbol{D}_N$ in descending order along its diagonal, $\Sigma_{1,N} = \mathrm{diag}(\mu_{1N}, \ldots, \mu_{K^*N})$, the columns of $\boldsymbol{U}_N$ contain the eigenvectors of $\boldsymbol{D}_N$ associated with the eigenvalues in Σ_N, $\boldsymbol{U}_N = \left(\boldsymbol{U}_{1,N}, \boldsymbol{U}_{2,N}\right)$, 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 if $z_i = z_j$ 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 $\tilde{\Sigma}_N = \mathrm{diag}(\tilde{\mu}_{1,N}, \ldots, \tilde{\mu}_{p,N})$ contains the first p eigenvalues of $\tilde{\boldsymbol{D}}_N$ in descending order. By Theorem 3.1, we can readily show that $\|\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N\| = O_P((p/T)^{1/2})$, ensuring that $\max_{1\leq\ell\leq N}\left|\tilde{\mu}_{\ell,N} - \mu_{\ell,N}\right| \leq \|\tilde{\boldsymbol{D}}_N - \boldsymbol{D}_N\| = O_P\left((p/T)^{1/2}\right)$, 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 $\sin\theta$ theorem in Davis and Kahan, 1970, Chapter VII in Bhatia, 1997, Proposition 2.1 in Rohe et al., 2011, Theorem 3 in Yu et al., 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 $\lambda_{K^*}(\boldsymbol{A}) = \lambda_{K^*}\left(\boldsymbol{\alpha}^0\boldsymbol{\alpha}^{0\top}\right) \geq c$ for some constant $c > 0$. Noting that AB and BA share the same set of nonzero eigenvalues, we have

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

It follows that $\lim_{N\to\infty}\mu_{K^*N} \geq c\min_{1\leq k\leq K^0}\tau_k > 0$ 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 $\lambda_{\mathcal{K}_N}(\tilde{\boldsymbol{D}}_N) \geq c_N$, 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 $(p/T)^{1/2}$. 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 $\tilde{\Sigma}_N = \mathrm{diag}(\tilde{\Sigma}_{1,N}, \tilde{\Sigma}_{2,N})$, 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 exists 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 and* $\max_{1\leq i\leq N}\left\|\beta_i^0\right\| = O(1)$. *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* $\max_{1\leq i\leq N}\sqrt{N}\left\|\tilde{u}_{1,i} - O_N u_{1,i}\right\| = O_P\left((p/T)^{1/2}(\ln T)^3\right)$.

An immediate implication of Theorem 4.2 is $\|\tilde{\boldsymbol{U}}_{1,N} - \boldsymbol{U}_{1,N}O_N\| = O_P\left((p/T)^{1/2}(\ln T)^3\right) = o_P(1)$, 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 $\boldsymbol{U}_{1,N} = \left(\boldsymbol{U}_{\cdot 1}, \ldots, \boldsymbol{U}_{\cdot \mathcal{K}_N}\right)$.[3] 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

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

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

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

which measures the variation of $\tilde{\mathcal{S}}_{i,l}(j)$ in the presence of a conjectured break point at m. We propose to adopt the following eigenvector-based SBSA to estimate $\mathcal{G}^0$.

Sequential Binary Segmentation Algorithm 2 (SBSA 2). SBSA 2 is essentially the same as with $\tilde{\boldsymbol{U}}_{1,N}$ $(N \times \mathcal{K}_N)$, $\tilde{\mathcal{S}}_{i,l}(j)$, $\bar{U}_{i,l}(j)$, $\tilde{V}_{i,l}(j)$ and $\tilde{S}_{i,l}(j,m)$ in place of $\tilde{\boldsymbol{\beta}}$ $(N \times p)$, $\mathcal{S}_{i,l}(j)$, $\bar{\beta}_{i,l}(j)$, $\hat{V}_{i,l}(j)$, and $\hat{S}_{i,l}(j,m)$ in SBSA, respectively.[4]

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 $\tilde{\mathcal{G}}\left(K^0\right) \equiv \{\tilde{G}_1, \tilde{G}_2, \ldots, \tilde{G}_{K^0}\}$ 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

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

Let

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

be the estimated number of groups and the estimated group structure, respectively. We will show that $P(\tilde{K} = K^0) \to 1$ and $P(\tilde{\mathcal{G}} = \mathcal{G}^0) \to 1$ 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 estimates of $\boldsymbol{\beta}$, $\boldsymbol{\mu}$, θ, and $\boldsymbol{\alpha}$. In particular, we denote the estimates of $\boldsymbol{\alpha}$ 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 subsection reports the asymptotic properties of $\tilde{\mathcal{G}}(K^0)$, $\tilde{K}$, $\tilde{\boldsymbol{\alpha}}$, and $\tilde{\theta}$.

[3] 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.

[4] For completeness, we provide the details of SBSA 2 in the online supplementary material.

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 criterion 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 convergence rate does not play a role here. Theorem 4.2 ensures that $\tilde{\boldsymbol{U}}_{1,N}$ contains all the individuals' group information that is required and it implies the uniform convergence 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 $\tilde{\mathcal{G}} = \tilde{\mathcal{G}}(\tilde{K})$ 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

Here we consider five data generating processes (hereafter DGPs). DGPs 1–3 specify linear panel data models while DGPs 4–5 consider a two-sided-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 for all DGPs except DGP 2. For DGP 2, we have $p = 10$ and cannot consider the case with $T = 10$ because we cannot obtain any reasonable preliminary estimates in this case. 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 (*Linear Panel With Diverging p*)**.** The data are generated as

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

where x_{it} is a $p \times 1$ vector with the jth element given by $x_{j,it} = 0.2\mu_i + e_{j,it}, j = 1, \ldots, p$, 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. And p takes 4, 6, and 8 for $T = 10$, 20, and 40, respectively. The true coefficients β_i can be classified into 3 groups with true group-specific parameter values given by

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left(\begin{bmatrix} -1 \\ -0.9 \\ \vdots \\ -1 + 0.1(p-1) \end{bmatrix}, \begin{bmatrix} 0.5 \\ 0.6 \\ \vdots \\ 0.5 + 0.1(p-1) \end{bmatrix}, \begin{bmatrix} 1 \\ 1.1 \\ \vdots \\ 1 + 0.1(p-1) \end{bmatrix} \right).$$

This DGP is designed to corroborate our asymptotic theory for the diverging p case and to evaluate its finite sample performance.

DGP 4 (*Two-sided-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 $e_{1,it}, e_{2,it}, \varepsilon_{it}, \mu_i$ 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 5 (*Dynamic One-sided 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 4. 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 4, 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 criterion in (2.8) to choose the number of groups. For DGPs 1–3, 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 4–5, 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 given 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 500 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 1–3 we also consider C-Lasso.[5] 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 DGPs 1 and 3, SBSA 2 slightly outperforms the C-Lasso and in DGP 2,

[5] 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 $K = 1, \dots, 5$ groups when $K^0 = 3$ and $K^{\max} = 5$.

	N	T	DGP 1					DGP 2					DGP 3				
			1	2	**3**	4	5	1	2	**3**	4	5	1	2	3	4	5
K-means on $\tilde{\beta}$	100	10	0	0	0.880	0.120	0						0	0.588	0.372	0.040	0
	100	20	0	0	0.874	0.126	0	0	0	0.928	0.066	0.006	0	0.102	0.876	0.022	0
	100	40	0	0	0.906	0.094	0	0	0	0.930	0.070	0	0	0.090	0.838	0.072	0
	200	10	0	0	0.872	0.116	0.012						0	0.278	0.678	0.044	0
	200	20	0	0	0.854	0.146	0	0	0	0.942	0.058	0	0	0.228	0.694	0.078	0
	200	40	0	0	0.870	0.128	0.002	0	0	0.932	0.068	0	0	0.198	0.722	0.080	0
K-means on eigenvectors	100	10	0	0.284	0.200	0.210	0.306						0	0.532	0.186	0.166	0.116
	100	20	0	0.060	0.414	0.348	0.178	0	0	0.978	0.022	0	0	0.212	0.384	0.170	0.234
	100	40	0	0	0.794	0.198	0.008	0	0	0.982	0.018	0	0	0.090	0.156	0.354	0.400
	200	10	0	0.146	0.234	0.274	0.346						0	0.700	0.160	0.118	0.022
	200	20	0	0.022	0.342	0.378	0.258	0	0	0.988	0.012	0	0	0.492	0.354	0.106	0.048
	200	40	0	0	0.734	0.262	0.004	0	0	0.994	0.006	0	0	0.224	0.290	0.264	0.222
C-Lasso	100	10	0	0	0.996	0.004	0						0	0.012	0.982	0.006	0
	100	20	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
	100	40	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
	200	10	0	0	0.994	0.006	0						0	0.020	0.974	0.006	0
	200	20	0	0	1	0	0	0	0	0.998	0.002	0	0	0.002	0.998	0	0
	200	40	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
SBSA 1	100	10	0	0.012	0.988	0	0						0	0.282	0.274	0.254	0.190
	100	20	0	0	1	0	0	0	0	0.102	0.898	0	0	0.608	0.346	0.046	0
	100	40	0	0	1	0	0	0	0	0.024	0.972	0.004	0	0.012	0.986	0.002	0
	200	10	0	0	0.998	0.002	0						0	0.246	0.368	0.268	0.118
	200	20	0	0	1	0	0	0	0	0.068	0.932	0	0	0.232	0.742	0.026	0
	200	40	0	0	1	0	0	0	0	0.006	0.994	0	0	0.006	0.994	0	0
SBSA 2	100	10	0	0	0.996	0.004	0						0	0.088	0.898	0.014	0
	100	20	0	0	1	0	0	0	0	0.994	0.006	0	0	0	1	0	0
	100	40	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
	200	10	0	0	1	0	0							0	1		0
	200	20	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
	200	40	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0

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 ratio 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 ratio of individuals falling into the right group. We show its average value across all replications in columns 4, 8 and 12 of Table 3 and in columns 4 and 8 of Table 4. Columns 5–7 and 9–11 (13–15) report the performance of the estimates of $\boldsymbol{\alpha}_{\cdot 2}^0 \equiv (\alpha_{1,2}^0, \dots, \alpha_{K^0,2}^0)^\top$, i.e., the second regressor's coefficient of all groups, in these two tables. 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 $\alpha_{k,2}^0$, $k = 1, \dots, K^0$, 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 on the post-classification ones.

We summarize some important findings from Tables 3 and 4. First, the ratio of correct classification generally increases with T for all classification methods under consideration for all DGPs but DGP 3. In DGP 3, the number of parameters (p) increases as T increases, which makes it more difficult to achieve correct classification with large T. In this case, the K-means based on the eigenvectors does not improve as T increases while the other methods still improve. 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 and bias and more accurate coverage probability 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 as T increases.

In general, the higher the correct classification ratio, the more accurate post-classification estimates we can obtain. When the correct classification ratio based on the same classification method improves (say, as T increases), we should

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

	N	T	DGP 4					DGP 5				
			1	2	**3**	4	5	1	2	**3**	4	5
K-means on $\tilde{\beta}$	100	10	0	0.050	0.882	0.068	0	0	0.072	0.702	0.224	0.002
	100	20	0	0.030	0.882	0.088	0	0	0.058	0.632	0.310	0
	100	40	0	0.002	0.946	0.052	0	0	0.032	0.666	0.294	0.008
	200	10	0	0	0.962	0.038	0	0	0.148	0.722	0.130	0
	200	20	0	0	0.898	0.102	0	0	0.122	0.710	0.168	0
	200	40	0	0	0.914	0.086	0	0	0.048	0.660	0.292	0
K-means on eigenvectors	100	10	0	0.012	0.734	0.248	0.006	0	0.178	0.708	0.102	0.012
	100	20	0	0	0.906	0.094	0	0	0.060	0.640	0.278	0.022
	100	40	0	0	0.886	0.114	0	0	0.030	0.644	0.312	0.014
	200	10	0	0.002	0.754	0.232	0.012	0	0.105	0.704	0.166	0.025
	200	20	0	0	0.862	0.138	0	0	0.080	0.688	0.224	0.008
	200	40	0	0	0.882	0.118	0	0	0.004	0.758	0.238	0
SBSA 1	100	10	0	0.120	0.634	0.230	0.016	0	0.058	0.816	0.120	0.006
	100	20	0	0	0.964	0.036	0	0	0	0.998	0.002	0
	100	40	0	0	0.998	0.002	0	0	0	1	0	0
	200	10	0	0	0.740	0.220	0.040	0	0.006	0.978	0.016	0
	200	20	0	0	0.988	0.012	0	0	0	1	0	0
	200	40	0	0	1	0	0	0	0	1	0	0
SBSA 2	100	10	0	0.004	0.996	0	0	0	0.002	0.996	0.002	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

also observe that the rmse and bias decrease. The story changes slightly when we compare different classification methods. Even when two classification methods perform *similarly* in terms of correct classification ratio, their performance can differ quite a bit in terms of rmse or bias because of the differences of the two classification methods. For example, for DGP 1 in Table 3, we observe that the K-means on $\tilde{\boldsymbol{\beta}}$ with $T = 40$ yields a higher correct classification ratio than SBSA 2 with $T = 10$ whereas the former's associated RMSE/bias is still larger than that of SBSA 2. Our simulations suggest that K-means sometimes generate very erratic classifications comparing with the true group structure, which results in a significant increase in the RMSE and bias. One potential reason behind this observation is that K-means selects initial centroids randomly and it is very much affected by such an initial choice. We think that this observation also helps to explain why the K-means methods often have a lower frequency of choosing the true number of groups than the SBSA method in Tables 1 and 2.

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 et al. (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 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.

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

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

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

	N	T	DGP 1				DGP 2				DGP 3			
			Correct ratio	Comparison criteria			Correct ratio	Comparison criteria			Correct ratio	Comparison criteria		
				RMSE	Bias	Coverage		RMSE	Bias	Coverage		RMSE	Bias	Coverage
Oracle	100	10	1	0.061	0.002	0.924					1	0.056	−0.004	0.944
	100	20	1	0.040	−0.002	0.917	1	0.043	0.000	0.925	1	0.041	−0.002	0.950
	100	40	1	0.027	0.000	0.958	1	0.028	0.003	0.954	1	0.029	0.002	0.952
	200	10	1	0.041	−0.001	0.931					1	0.039	−0.002	0.962
	200	20	1	0.027	−0.002	0.950	1	0.028	−0.000	0.944	1	0.031	0.003	0.943
	200	40	1	0.019	0.001	0.946	1	0.020	0.001	0.950	1	0.020	0.001	0.948
K-means on $\tilde{\beta}$	100	10	0.884	0.301	−0.081	0.764					0.747	0.312	−0.176	0.528
	100	20	0.914	0.293	−0.106	0.787	0.970	0.171	−0.031	0.883	0.932	0.178	−0.053	0.850
	100	40	0.960	0.214	−0.055	0.886	0.962	0.195	−0.028	0.900	0.936	0.181	−0.052	0.866
	200	10	0.894	0.247	−0.072	0.760					0.806	0.238	−0.104	0.702
	200	20	0.923	0.272	−0.091	0.834	0.979	0.128	−0.023	0.920	0.865	0.243	−0.103	0.690
	200	40	0.945	0.227	−0.067	0.854	0.967	0.270	−0.029	0.894	0.932	0.170	−0.055	0.864
K-means on eigenvectors	100	10	0.702	0.537	−0.243	0.334					0.641	0.447	−0.176	0.292
	100	20	0.778	0.383	−0.191	0.491	0.978	0.126	−0.007	0.881	0.582	0.532	−0.121	0.200
	100	40	0.902	0.278	−0.106	0.794	0.993	0.123	0.006	0.943	0.529	0.602	−0.078	0.078
	200	10	0.695	0.544	−0.254	0.316					0.657	0.379	−0.157	0.322
	200	20	0.775	0.376	−0.171	0.430	0.986	0.065	−0.008	0.905	0.637	0.452	−0.103	0.250
	200	40	0.873	0.316	−0.142	0.735	0.993	0.107	−0.003	0.940	0.585	0.502	−0.066	0.122
C-Lasso	100	10	0.941	0.076	−0.015	0.861					0.935	0.069	−0.015	0.866
	100	20	0.986	0.044	−0.006	0.900	1	0.043	0.000	0.925	0.995	0.043	−0.004	0.942
	100	40	0.999	0.027	−0.000	0.957	1	0.028	0.003	0.954	1	0.029	0.002	0.952
	200	10	0.942	0.053	−0.018	0.829					0.929	0.050	−0.014	0.854
	200	20	0.986	0.028	−0.005	0.938	1	0.028	−0.000	0.944	0.995	0.035	0.002	0.927
	200	40	0.999	0.019	0.001	0.937	1	0.020	0.001	0.950	1	0.020	0.001	0.948
SBSA 1	100	10	0.930	0.089	0.006	0.851					0.737	0.261	−0.152	0.442
	100	20	0.984	0.043	−0.003	0.901	0.780	0.518	−0.045	0.327	0.887	0.093	−0.001	0.614
	100	40	0.999	0.027	0.000	0.959	0.851	0.287	−0.024	0.314	0.953	0.045	0.001	0.806
	200	10	0.932	0.052	0.003	0.856					0.754	0.205	−0.095	0.397
	200	20	0.985	0.029	−0.001	0.934	0.772	0.481	−0.046	0.314	0.884	0.085	−0.005	0.560
	200	40	0.999	0.019	0.001	0.943	0.853	0.224	−0.026	0.294	0.953	0.033	0.002	0.734
SBSA 2	100	10	0.931	0.077	0.005	0.860					0.911	0.066	0.003	0.848
	100	20	0.985	0.043	−0.003	0.911	0.991	0.047	0.001	0.894	0.989	0.042	−0.003	0.945
	100	40	0.998	0.027	−0.000	0.958	1	0.028	0.003	0.954	1	0.029	0.002	0.952
	200	10	0.930	0.051	0.005	0.862					0.911	0.044	0.003	0.831
	200	20	0.984	0.030	−0.002	0.925	0.992	0.032	0.001	0.911	0.990	0.032	0.003	0.923
	200	40	0.999	0.019	0.001	0.942	1	0.020	0.001	0.950	1	0.020	0.001	0.948

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.[6][7] 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, $\beta^0_{1i} = \sum_{k=1}^{K^0} \alpha^0_k \cdot \mathbf{1}\left\{i \in G^0_k\right\}$. 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, the 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

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

[7] Note that the time trend is nonstationary and our asymptotic theory does not apply to this case directly. But one can readily modify our asymptotic analysis to allow for the time trend by permitting different convergence rates for different parameter estimators.

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

	N	T	DGP 4				DGP 5			
			Correct ratio	Comparison criteria			Correct ratio	Comparison criteria		
				RMSE	Bias	Coverage		RMSE	Bias	Coverage
Oracle	100	10	1	0.089	-0.001	0.935	1	0.073	0.004	0.942
	100	20	1	0.059	0.002	0.941	1	0.046	0.001	0.953
	100	40	1	0.044	0.004	0.958	1	0.036	0.002	0.937
	200	10	1	0.066	0.002	0.932	1	0.056	−0.003	0.928
	200	20	1	0.043	0.004	0.927	1	0.041	−0.004	0.932
	200	40	1	0.031	0.000	0.943	1	0.029	−0.002	0.945
K-means on $\tilde{\beta}$	100	10	0.930	0.231	−0.029	0.830	0.882	0.268	−0.053	0.628
	100	20	0.967	0.148	−0.013	0.810	0.915	0.272	−0.079	0.672
	100	40	0.982	0.188	−0.026	0.852	0.923	0.278	−0.095	0.642
	200	10	0.934	0.179	−0.005	0.795	0.864	0.233	−0.068	0.620
	200	20	0.967	0.167	−0.016	0.802	0.914	0.219	−0.059	0.633
	200	40	0.985	0.175	−0.030	0.869	0.927	0.204	−0.054	0.684
K-means on eigen-vectors	100	10	0.770	0.338	0.052	0.445	0.834	0.317	−0.084	0.531
	100	20	0.927	0.276	0.029	0.769	0.912	0.286	−0.042	0.654
	100	40	0.918	0.255	−0.007	0.728	0.907	0.269	−0.059	0.627
	200	10	0.751	0.351	0.041	0.396	0.806	0.326	−0.085	0.548
	200	20	0.906	0.288	0.023	0.714	0.825	0.294	−0.053	0.602
	200	40	0.917	0.228	0.014	0.745	0.916	0.231	−0.048	0.671
SBSA 1	100	10	0.882	0.173	0.059	0.538	0.876	0.133	0.021	0.653
	100	20	0.956	0.089	0.028	0.837	0.948	0.072	0.001	0.847
	100	40	0.992	0.046	-0.001	0.932	0.984	0.039	0.003	0.917
	200	10	0.884	0.176	0.088	0.463	0.877	0.125	0.022	0.592
	200	20	0.963	0.078	0.037	0.756	0.955	0.056	-0.002	0.767
	200	40	0.992	0.033	0.001	0.925	0.986	0.032	−0.003	0.933
SBSA 2	100	10	0.931	0.107	0.033	0.861	0.924	0.088	0.007	0.875
	100	20	0.987	0.064	0.011	0.928	0.971	0.052	0.006	0.941
	100	40	0.997	0.045	0.003	0.947	0.993	0.037	-0.001	0.933
	200	10	0.938	0.100	0.039	0.824	0.927	0.067	0.016	0.878
	200	20	0.985	0.049	0.007	0.916	0.976	0.045	0.003	0.908
	200	40	0.998	0.032	0.003	0.942	0.997	0.030	−0.000	0.940

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

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 with y_{it}, the ratio of safe assets 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 $N^{-1} \sum_{i=1}^{N} T_i \approx 12.3$. 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.

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. Fig. 2 reports the scatter diagram for the preliminary estimates of β_{1i}, viz, the slope coefficients of log(FA) and log(NCI) along with the individual estimated group identity. Even though one cannot well separate individuals in a group from those in the other groups simply via eyeballing, the patterns for the SBSA2-based classification are clear.

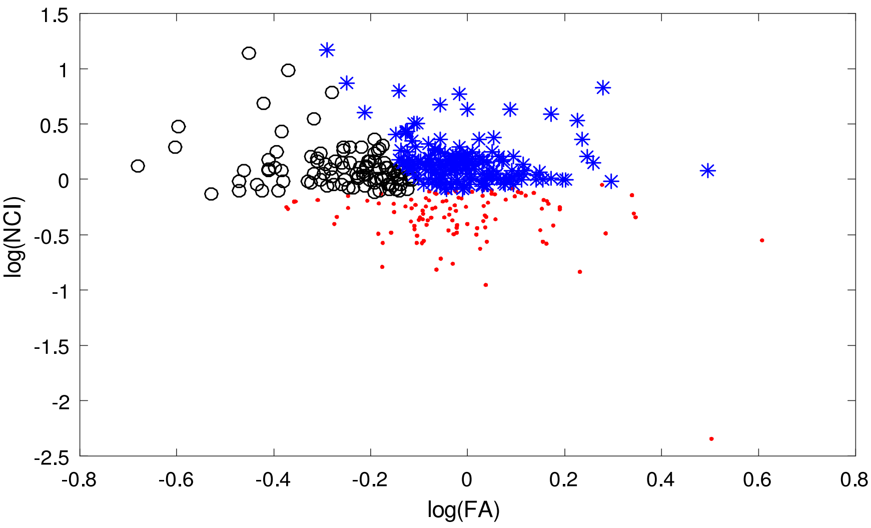

Fig. 2. Scatter plot for the preliminary estimates of β_{1i}'s in the application. The x-axis and y-axis mark the estimates of the slope coefficients of log financial assets (log(FA)) and log non-capital income (log(NCI)), respectively. By applying SBSA 2, we identify 3 groups. Groups 1, 2 and 3 are marked with red dots, black circles, and blue stars, respectively.

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***	−0.055***	−0.223***	−0.048***	−0.129***
	(0.005)	(0.009)	(0.009)	(0.008)	(0.011)
log(NCI)	0.035***	−0.255***	0.056***	0.091***	−0.006*
	(0.012)	(0.024)	(0.018)	(0.016)	(0.004)
AEX premium	0.008	−0.007			−0.039**
	(0.023)	(0.022)			(0.017)
Time (t)	0.024***	0.020***			−0.013***
	(0.001)	(0.001)			(0.002)
Retirement dummy	0.079***	0.065***			
	(0.021)	(0.020)			
σ^2	0.310***	0.290***			
	(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 the regression results drawn from AS for comparison purpose. Standard errors are in parentheses.
*Significance at 10% level.
**Significance at 5% level.
***Significance at 1% level.

First, the estimate of the slope coefficients of log(NCI) tends to be negative for individuals in Group 1 and positive for individuals in the other two groups. Second, even though log(NCI) tends to have positive effects on the ratio of safe assets for the individuals in both Groups 2 and 3, the effect of log(FA) in Group 2 tends to negative and stronger than that in Group 3.

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

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^{***}	-0.240^{***}	-0.055^{***}	-0.040^{***}	-0.224^{***}	-0.028^{***}
	(0.008)	(0.009)	(0.007)	(0.008)	(0.007)	(0.005)
log(NCI)	-0.304^{***}	0.027	0.068^{***}	-0.414^{***}	0.031^{**}	0.028^{**}
	(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^{***}			0.022^{***}	
		(0.001)			(0.001)	
Retirement dummy		0.069^{***}			0.052^{***}	
		(0.018)			(0.015)	
σ^2		0.290^{***}			0.266^{***}	
		(0.004)			(0.003)	

**Significance at 5% level.

***Significance at 1% level.

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 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 reported 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 a 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 the 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.

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 by 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 information criteria 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 exhibit oracle efficiency. 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 $\sqrt{T}$-consistent, which enables us to conduct the SBSA as in the current paper to detect possible group 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 (2018) 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 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 unit-root-type regressors in our model as in Huang et al. (2020). 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 (2021). We leave these topics for future research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.04.003.

References

Abrevaya, J., Shen, S., 2014. Estimation of censored panel data models with slope heterogeneity. J. Appl. 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. J. Econom. 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. J. Appl. Econometrics 31, 163–191.

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

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

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

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. In: Blundell, R., Newey, W.K., Persson, T. (Eds.), Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. 3. Cambridge University Press, New York, pp. 45–74.

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.

Cocco, J.F., Gomes, F.J., Maenhout, P.J., 2005. Consumption and portfolio choice over the life cycle. Rev. Financ. Stud. 18, 491–533.

Curcuru, S., Heaton, J., Lucas, D., Moore, D., 2004. Heterogeneity and portfolio choice: theory and evidence. Handb. Financ. Econom. 1, 337–382.

Davis, C., Kahan, W.M., 1970. The rotation of eigenvectors by a perturbation: III. SIAM J. Numer. Anal. 7, 1–46.

Dhaene, G., Jochmans, K., 2015. Split-panel jackknife estimation of fixed-effect models. Rev. Econom. Stud. 82, 991–1030.

Fan, J., Lv, J., Qi, L., 2011. Sparse high dimensional models in economics. Annu. Rev. Econ. 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.

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. J. Amer. Statist. Assoc. 92, 455–465.

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

Huang, W., Jin, S., Su, L., 2020. Panel cointegration with latent group structures and an application to the PPP theory. Econometric Theory forthcoming.

Ke, Z.T., Fan, J., Wu, Y., 2015. Homogeneity pursuit. J. Amer. Statist. Assoc. 110, 175–194.

Ke, Y., Li, J., Zhang, W., 2016. Structure identification in panel data analysis. Ann. Statist. 44, 1193–1233.

Lam, C., Fan, J., 2008. Profile-kernel likelihood inference with diverging number of parameters. Ann. Statist. 36, 2232–2260.

Lin, C.-C., Ng, S., 2012. Estimation of panel data models with parameter heterogeneity when group membership is unknown. J. Econom. 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. Quant. Econ. 8, 729–760.

Ma, S., Huang, J., 2017. A concave pairwise fusion approach to subgroup analysis. J. Amer. Statist. Assoc. 112, 410–423.

Okui, R., Wang, W., 2021. Heterogeneous structural breaks in panel data models. J. Econometrics 220, 447–473.

Pesaran, M.H., Shin, Y., Smith, R.P., 1999. Pooled mean group estimation of dynamic heterogeneous panels. J. Amer. Statist. Assoc. 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. Rev. Financ. Stud. 20, 83–124.

Qian, J., Su, L., 2016. Shrinkage estimation of regression models with multiple structural changes. Econometric Theory 32, 1376–1433.

Radchenko, P., Mukherjee, G., 2017. Convex clustering via l_1 fusion penalization. J. R. Stat. Soc. Ser. B Stat. Methodol. 79, 1527–1546.

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

Samuelson, P.A., 1969. Lifetime portfolio selection by dynamic stochastic programming. Rev. Econ. Stat. 23, 9–246.

Sarafidis, V., Weber, N., 2015. A partially heterogeneous framework for analyzing panel data. Oxf. Bull. Econ. Stat. 77, 274–296.

Shen, J., He, X., 2015. Inference for subgroup analysis with a structured logistic-normal mixture model. J. Amer. Statist. Assoc. 110, 303–312.

Shen, X., Huang, H.-C., 2010. Grouping pursuit through a regularization solution surface. J. Amer. Statist. Assoc. 105, 727–739.

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

Su, L., Ju, G., 2018. Identifying latent grouped patterns in panel data models with interactive fixed effects. J. Econometrics 206, 554–573.

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. Econom. Suppl. Mater. 84, http://dx.doi.org/10.3982/ECTA12560.

Su, L., Wang, X., Jin, S., 2019. Sieve estimation of time-varying panel data models with latent structures. J. Bus. Econom. Statist. 37, 334–349.

Subramanian, A., Wei, S.-J., 2007. The WTO promotes trade, strongly but unevenly. J. Int. Econ. 72, 151–175.

von Luxburg, U., 2007. A tutorial on spectral clustering. Stat. Comput. 17, 395–416.

Wang, W., Phillips, P.C.B., Su, L., 2018. Homogeneity pursuit in panel data models: theory and applications. J. Appl. Econometrics 33, 797–815.

Wang, W., Phillips, P.C.B., Su, L., 2019. The heterogeneous effects of the minimum wage on employment across states. Econom. Lett. 174, 179–185.

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