Identifying Latent Group Structures in Nonlinear Panels*  

Wuyi Wang\textsuperscript{a} and Liangjun Su\textsuperscript{b}  
\textsuperscript{a} Institute for Economic and Social Research, Jinan University  
\textsuperscript{b} School of Economics, Singapore Management University  

January 9, 2019

Abstract

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

JEL Classification: C33, C38, C51.

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

*The authors thank the 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 (SETA 2017, Beijing), and the Advance in Econometrics Conference (2017, Shanghai), and the seminar participants in the Department of Economics at University of York, the Department of Economics at Humboldt-Universitat Zu Berlin, Germany, and the Erasmus School of Economics at Erasmus Universiteit Rotterdam, for their valuable comments. Su acknowledges support from the Singapore Ministry of Education for Academic Research Fund (AcRF) under the Tier-2 grant number MOE2012-T2-2-021 and the funding support provided by the Lee Kong Chian Fund for Excellence. 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.
1 Introduction

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

In a panel structure model, there exists a subset of parameters that are heterogeneous across groups but homogeneous within a group, and neither the number of groups nor individuals’ group membership is known. There are many motivating examples for such a model. In macroeconomics, Phillips and Sul (2007) study the hypothesis of convergence clubs where countries belonging to different groups behave differently; in financial markets, stocks in the same sector share some similar characteristics and behave similarly (Ke, Fan, and Wu (2015)). In 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, Lv, and Qi (2011); Bester and Hansen (2016)); in international trade, GATT/WTO has uneven impacts on different groups of country-pairs (Subramanian and Wei (2007)). All these examples motivate the use of panel structure models.

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), Ando and Bai (2016). The second popular approach is based on the classifier-Lasso (C-Lasso) that has been recently proposed by Su, Shi, and Phillips (2016a, SSP hereafter) and extended in Su and Ju (2018), Su, Wang, and Jin (2019), and Wang, Phillips, and Su (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, Phillips, and Su (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, Li, and Zhang (2016, KLZ hereafter) borrow the idea of binary segmentation in the structural change literature (e.g., Bai (1997)) and apply it to identify the unobserved group structures in linear panel data models with interactive fixed effects. Let $N$ denote the number of cross sectional units and $p$ the dimension of a parameter vector $\beta_i$ that is associated with individual $i$. Let $B = (\beta_1^T, \ldots, \beta_N^T)^T$. KLZ assume that the number of distinct elements in the $Np$-vector $B$ is given by a finite number, say $N + 1$ in their notation. Based on consistent preliminary estimates $\tilde{B}$ of $B$, they order the elements of $\tilde{B}$ 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 $B$. Apparently, the setup in KLZ is quite different from the general setup in econometrics where the parameters of interest, $\beta_i$ as a whole vector, are assumed to be heterogeneous across groups but homogeneous within a group.

Following the lead of Bai (1997) and KLZ, we propose to apply the sequential binary segmentation algorithm (SBSA) to identify the latent group structure on parameter vectors in nonlinear panel data models. In comparison with KLZ, our method is different from theirs in three important ways. First, KLZ consider the classification of scalar coefficients but we consider the classification of parameter vectors. In KLZ’s case, there is a natural ordering for their preliminary estimates and they can draw support from the structural change literature where parameters of interest are ordered naturally along the time dimension. In our case, there is no natural order for the estimates of parameter vectors, and fortunately, inspired by the CART-split criterion (Breiman, Friedman, Stone, and Olshen (1984)), we are able to propose a variant of binary segmentation algorithm to classify the vectors. Second, KLZ consider the linear panel data models with interactive fixed effects. They obtain their preliminary estimates by using an EM algorithm and then conduct the binary segmentation based on the ordered preliminary estimates. In contrast, we consider general nonlinear panel data models that 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 asymp-
totic 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 \textit{a priori}, there is no tuning parameter involved in our procedure and the K-means algorithm as well, and one tuning parameter is involved in the C-Lasso procedure. Third, SSP’s C-Lasso may leave some individuals unclassified and one has to classify some unclassified individuals after the algorithm based on some distance measure. Like the K-means algorithm, our binary segmentation algorithm forces all individuals to be classified into one of the groups. As SSP argue, leaving some individuals unclassified is not necessarily a bad thing. We also find through our simulations that the preliminary estimates based on some realizations can be rather abnormal when the time dimension $T$ in the panel is not large. In this case, including such abnormal estimates in the algorithm can significantly deteriorate the classification performance. Fourth, in some sense, our method can be regarded as a universal method and it works for all panel structure models as long as one can obtain preliminary consistent estimates. The model can be nonstationary panels or panel data models with interactive fixed effects.

In addition, we also allow the presence of common parameters across all individuals. This corresponds to the mixed panel structure model mentioned in SSP (Section 2.7). It is useful when economic theory suggests that some regressors’ coefficients are identical across individuals (e.g., Pesaran, Shin, and Smith (1999)) 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.\footnote{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.}

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

In the application, we study how individuals’ portfolio choices are affected by financial assets, non-capital income, retirement status and other factors. Among them, financial assets and non-capital income are modeled to have heterogeneous responses for different individuals. The response variable is the safe asset ratio, which is left censored at 0 and right censored at 1. We use data from the De Nederlandsche Bank (DNB) panel survey. By using the method proposed here, we are able to identify three latent groups. The first group of individuals 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.

\textbf{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, $\lambda_{\text{max}}(A)$, $\lambda_{\text{min}}(A)$, and $\lambda_j(A)$ denote its largest, smallest, and $j$th largest eigenvalues, respectively. $I_p$ and $0_{p \times 1}$ denote the $p \times p$ identity matrix and $p \times 1$ vector of zeros, respectively. $\mathbf{1}\{\cdot\}$ denotes the indicator function. The operators $\xrightarrow{D}$ and $\xrightarrow{P}$ denote convergence in distribution and in probability, respectively.
2 The model and the estimators

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

2.1 The panel structure model and examples

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

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

(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., \( y_{i,t-1} \)), \( \varepsilon_{it} \) is the idiosyncratic shock, \( \mu_i \) is an \( r \times 1 \) vector of nuisance parameters (e.g., the fixed effects), \( \theta \) is a \( q \times 1 \) vector of parameters that are common across individuals, and \( \beta_i \) is a \( p \times 1 \) vector of parameters whose true values exhibit a group pattern of the general form

\[ \beta_i^0 = \sum_{k=1}^{K_0} \alpha_k^0 \cdot 1 \{ i \in G_k^0 \}. \]

Here \( \alpha_k^0 \neq \alpha_l^0 \) for any \( k \neq l \) and \( G^0 = \{ G_1^0, \ldots, G_{K_0}^0 \} \) 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 \( G^0 \) 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 \( \beta = (\beta_1, \ldots, \beta_N)^\top \), \( \alpha = (\alpha_1, \ldots, \alpha_{K_0})^\top \), and \( \mu = (\mu_1, \ldots, \mu_N)^\top \). The true values of \( \beta, \alpha, \mu, \) and \( \theta \) are denoted by \( \beta^0, \alpha^0, \mu^0, \) and \( \theta^0 \), respectively. Without any information about the group structure, we propose to minimize the following objective function

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

(2.2)

When the likelihood function is correctly specified, by minimizing the above function we obtain the maximum likelihood estimates (MLEs) \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_N)^\top \), \( \hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{K_0})^\top \), and \( \hat{\theta} \) of \( \beta, \mu, \) and \( \theta \), 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

\[ y_{it} = x_{it}^\top \beta_i^0 + \mu_i^0 + \varepsilon_{it}, \]  

(2.3)

where \( \mu_i \) is the scalar fixed effect so that \( r = 1 \), \( \beta_i \), \( x_{it} \), and \( \varepsilon_{it} \) are defined as above, and the model does not contain any common parameter of interest so that \( \theta \) 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 \( w_{it} = (y_{it}, x_{it}^\top) \).
Following Pesaran et al. (1999), we can consider a mixed linear panel data model that contains both homogeneous and heterogeneous slope coefficients:

\[ y_{it} = x_{1,it}^T \beta_i^0 + x_{2,it}^T \theta_i^0 + \mu_i + \varepsilon_{it}, \]

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

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

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

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

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

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 \( I^L_{it} = 1\{y_{it} = L\} \) and \( I^R_{it} = 1\{y_{it} = R\} \). We consider four cases.

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

\[ y_{it}^* = x_{it}^T \beta_i^0 + \mu_i + \varepsilon_{it}, \]

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

\[
-\varphi(w_{it}; \beta_i, \mu_i, \sigma^2) = I^L_{it} \ln \Phi \left( \frac{(y_{it} - x_{it}^T \beta_i - \mu_i)/\sigma}{\sigma} \right) + I^R_{it} \ln \left( 1 - \Phi \left( \frac{(y_{it} - x_{it}^T \beta_i - \mu_i)/\sigma}{\sigma} \right) \right) \\
+ (1 - I^L_{it} - I^R_{it}) \ln \left[ \phi \left( \frac{(y_{it} - x_{it}^T \beta_i - \mu_i)/\sigma}{\sigma} \right) /\sigma \right],
\]

(2.4)

where \( \phi \) and \( \Phi \) 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:

\[ y_{it}^* = x_{1,it}^T \beta_i^0 + x_{2,it}^T \theta_i^0 + \mu_i + \varepsilon_{it}, \]

where \( \theta = (\sigma^2, \theta_2^T)^T \) and \( \theta_2 \) is a \((q - 1)\)-vector. The QMLE objective function follows directly from (2.4) with \( y_{it} - x_{1,it}^T \beta_i - \mu_i \) being replaced by \( y_{it} - x_{1,it}^T \beta_i - x_{2,it}^T \theta - \mu_i \).
example 2.3

The main interest of this paper is to identify the group structure about the number of groups and all individuals’ group membership.

2.2 Sequential binary segmentation algorithm

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

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

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

(i). The model is \( y_{it} = \mathbf{1}\{x_{it}^\top \beta_0^0 + \mu_i^0 - \varepsilon_{it} \geq 0\} \), where \( x_{it}, \beta_i, \) and \( \mu_i \) are defined as in Example 2.1 and \( \varepsilon_{it} \)'s are i.i.d. \( N(0, 1) \). So in this case, \(-\varphi(w_{it}; \beta_i, \mu_i) = y_{it} \ln \Phi(y_{it} - x_{it}^\top \beta_i - \mu_i) + (1 - y_{it}) \ln [1 - \Phi(y_{it} - x_{it}^\top \beta_i - \mu_i)] \).

(ii). The model is \( y_{it} = \mathbf{1}\{x_{1,it}^\top \beta_0^0 + x_{2,it}^\top \theta^0 + \mu_0^0 - \varepsilon_{it} \geq 0\} \). 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 \( G^0 \), 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 \( \tilde{\beta} \equiv (\tilde{\beta}_1, \ldots, \tilde{\beta}_N)^\top \) as

\[
\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_p),
\]

where \( \tilde{\beta}_j \) denotes the \( j \)-th column of \( \tilde{\beta} \) for \( j = 1, \ldots, p \). Let \( \beta_{i,j}^0, \alpha_{k,j}^0 \) and \( \tilde{\beta}_{i,j} \) denote the \( j \)-th element of \( \beta_i^0, \alpha_k^0 \) and \( \tilde{\beta}_i \), respectively, for \( j = 1, \ldots, p \). We sort the \( N \) elements of \( \tilde{\beta}_j \) in ascending order and denote the order statistics by

\[
\tilde{\beta}_{\pi_j(1), j} \leq \tilde{\beta}_{\pi_j(2), j} \leq \cdots \leq \tilde{\beta}_{\pi_j(N), j},
\]

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
\( S_{1,N} (j) = \{ \tilde{\beta}_{\pi_j(1),j}, \tilde{\beta}_{\pi_j(2),j}, \ldots, \tilde{\beta}_{\pi_j(N),j} \} \) 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 \( S_{1,N_1} (j), S_{N_1+1,N_1+N_2} (j), \) and \( S_{N_1+N_2+1,N} (j) \) 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 \( j \in \{1, \ldots, p\} \), \( \alpha_{1,j}^0, \ldots, \), and \( \alpha_{K^0,j}^0 \) are not all distinct from each other and \( K^0 \) is typically unknown. For this reason, we have to allow the possibility that \( \{ \alpha_{k,j}^0, k = 1, \ldots, K^0 \} \) are not all distinct from each other for all \( j \) and the possibility that \( \alpha_{1,j}^0 = \cdots = \alpha_{K^0,j}^0 \) for some \( j \). We achieve the identification of all \( K^0 \) groups based on the key observation that the sample variance of the subsample \( S_{i,l} (j) \) behaves quite differently depending on whether \( \beta_{\pi_j(i),j}^0 \) is the same as \( \beta_{\pi_j(l),j}^0 \). If \( \beta_{\pi_j(i),j}^0 = \beta_{\pi_j(i+1),j}^0 = \cdots = \beta_{\pi_j(l),j}^0 \), then the sample variance of \( S_{i,l} (j) \) 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 \( \beta_{\pi_j(i),j}^0 < \beta_{\pi_j(l),j}^0 \), then the sample variance of \( S_{i,l} (j) \) 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
\[
\tilde{\beta}_{i,l} (j) = \frac{1}{l-i+1} \sum_{i'=i}^{l} \tilde{\beta}_{\pi_j(i'),j} \quad \text{and} \quad \tilde{V}_{i,l} (j) = \frac{1}{l-i} \sum_{i'=i}^{l} [\tilde{\beta}_{\pi_j(i'),j} - \tilde{\beta}_{i,l} (j)]^2
\]
denote the sample mean and variance of the subsample \( S_{i,l} (j) \), respectively. Let \( \tilde{\sigma}_{i,l}^2 (j) \) denote a consistent estimator of the asymptotic variance \( \operatorname{Var}(\sqrt{T}\tilde{\beta}_{\pi_j(i),j}) \). Let
\[
\tilde{\sigma}_{i,l}^2 (j) = \frac{1}{l-i+1} \sum_{i'=i}^{l} \tilde{\sigma}_{i'}^2 (j)
\]
define
\[
\tilde{S}_{i,l} (j,m) = \frac{1}{l-i+1} \left\{ \sum_{i'=i}^{m} [\tilde{\beta}_{\pi_j(i'),j} - \tilde{\beta}_{i,m} (j)]^2 + \sum_{i'=m+1}^{l} [\tilde{\beta}_{\pi_j(i'),j} - \tilde{\beta}_{m+1,l} (j)]^2 \right\}, \tag{2.6}
\]
which measures the variation in \( S_{i,l} (j) \) 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 \( 1 \leq K^0 \leq K^{\max} \). Let \( K \) denote a generic number of groups. We propose to adopt the following SBSA to estimate \( K^0 \).

**Sequential Binary Segmentation Algorithm 1 (SBSA)**

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

\(^2\) A major difference between our algorithm and that of KLZ is that KLZ specify a tuning parameter \( \delta \) that is compared with something similar to our \( S_{i,N} (j,m) \) to determine when one should stop the algorithm. Even though they propose to use the BIC to choose \( \delta \), 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 \( \delta \) in KLZ but our result is insensitive to its choice.
2. When $K = 2$, let $\hat{j}_1 = \arg \max_{1 \leq j \leq p} \hat{V}_{1,N}(j)$. Given $\hat{j}_1$, we solve the following minimization problem
\[
\hat{m}_1 = \arg \min_{1 \leq m < N} \hat{S}_{1,N}(\hat{j}_1, m),
\]
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 $K \geq 3$, we use $\hat{m}_1, \ldots, \hat{m}_{K-2}$ denote the break points detected in the previous steps such that $\hat{m}_1 < \cdots < \hat{m}_{K-2}$ perhaps after relabeling the $K - 2$ break points that have been detected so far. Define
\[
\hat{j}_{K-1} = \arg \max_{1 \leq j \leq p} \sum_{k=1}^{K-1} \hat{V}_{\hat{m}_{k-1}+1,\hat{m}_k}(j),
\]
\[
\hat{m}_{K-1}(k) = \arg \min_{\hat{m}_{k-1}+1 \leq m < \hat{m}_k} \hat{S}_{\hat{m}_{k-1}+1,\hat{m}_k}(\hat{j}_{K-1}, m)
\]
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 $k = 1, \ldots, K - 1$,
\[
\hat{S}_{K-1}(k) = \sum_{i \in \hat{G}_{k1}(K-1)} \left[ \tilde{\beta}_{i,\hat{j}_{K-1}} - \tilde{\beta}_{\hat{G}_{k1}(K-1)}(\hat{j}_{K-1}) \right]^2 + \sum_{i \in \hat{G}_{k2}(K-1)} \left[ \tilde{\beta}_{i,\hat{j}_{K-1}} - \tilde{\beta}_{\hat{G}_{k2}(K-1)}(\hat{j}_{K-1}) \right]^2
\]
\[
+ \sum_{1 \leq l \leq K-1, l \neq k} \sum_{i \in \hat{G}_l(K-1)} \left[ \tilde{\beta}_{i,\hat{j}_{K-1}} - \tilde{\beta}_{\hat{G}_l(K-1)}(\hat{j}_{K-1}) \right]^2,
\]
where, e.g., $\tilde{\beta}_{\hat{G}_{k1}(K-1)}(\hat{j}_{K-1}) = |\hat{G}_{k1}(K - 1)|^{-1} \sum_{i \in \hat{G}_{k1}(K - 1)} \tilde{\beta}_{i,\hat{j}_{K-1}}$. 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
\[
\hat{k} = \arg \min_{1 \leq k \leq K-1} \hat{S}_{K-1}(k),
\]
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 $\hat{G}(K^0) = \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{K^0}\}$ 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 \( \hat{G}(K) \equiv \{ \hat{G}_1(K), \hat{G}_2(K), \ldots, \hat{G}_K(K) \} \). Given the estimated group structure \( \hat{G}(K) \) for \( K \in [1, K_{\text{max}}] \), we propose to estimate the model parameters by minimizing

\[
L_{NT}(\beta, \mu, \theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \varphi(w_{it}; \beta_i, \mu_i, \theta)
\]

s.t. \( \beta_i = \alpha_k \) for \( i \in \hat{G}_k(K) \) and \( k = 1, \ldots, K \). \hspace{1cm} (2.7)

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

\[
\text{IC}_1(K) = 2L_{NT}(\hat{\beta}(K), \hat{\mu}(K), \hat{\theta}(K)) + pK \cdot \rho_{NT},
\]

where \( \rho_{NT} \) is a function of \( (N, T) \). It plays the role of \( \ln(NT)/(NT) \) in the use of BIC in the panel setup. Let

\[
\hat{K} \equiv \arg \min_{1 \leq K \leq K_{\text{max}}} \text{IC}_1(K) \quad \text{and} \quad \hat{\hat{G}} \equiv \hat{G}(\hat{K}) \equiv \{ \hat{G}_1(\hat{K}), \hat{G}_2(\hat{K}), \ldots, \hat{G}_K(\hat{K}) \}
\] \hspace{1cm} (2.9)

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

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

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

\[
\hat{\beta} \equiv \hat{\beta}(\hat{K}) = (\hat{\beta}_1(\hat{K}), \ldots, \hat{\beta}_N(\hat{K}))^\top, \quad \hat{\mu} \equiv \hat{\mu}(\hat{K}) = (\hat{\mu}_1(\hat{K}), \ldots, \hat{\mu}_N(\hat{K}))^\top,
\]

\[
\hat{\alpha} \equiv \hat{\alpha}(\hat{K}) = (\hat{\alpha}_1(\hat{K}), \ldots, \hat{\alpha}_K(\hat{K}))^\top, \quad \hat{\theta} \equiv \hat{\theta}(\hat{K}).
\]

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

\[
Q_{NT}(\alpha, \theta) = \frac{1}{NT} \sum_{k=1}^{\hat{K}} \sum_{i \in \hat{G}_k(\hat{K})} \sum_{t=1}^{T} \varphi(w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k, \theta), \theta),
\]

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 \( \hat{\alpha} \) 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 \( \gamma_i = (\beta_i^T, \mu_i^T)^T, \varsigma_i = (\gamma_i^T, \theta^T)^T, \gamma_i^0 = (\beta_i^0, \mu_i^0)^T, \) and \( \varsigma_i^0 = (\gamma_i^0, \theta^0)^T \). 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

\[
Q_{NT}(\theta) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \phi(w_{it}; \gamma_i(\theta)),
\]

where \( \gamma_i(\theta) = \arg \min_{\gamma_i} \frac{1}{T} \sum_{t=1}^{T} \phi(w_{it}; \gamma_i, \theta) \).

Let \( \hat{\theta} = \arg \min_{\theta} Q_{NT}(\theta) \) and \( \gamma_i = \gamma_i(\hat{\theta}) = (\beta_i^T, \mu_i^T)^T \).

Let

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

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

Let \( Z(w_{it}; \gamma_i, \theta) = \partial \phi(w_{it}; \gamma_i, \theta) / \partial \gamma_i \) and \( W(w_{it}; \gamma_i, \theta) = \partial \phi(w_{it}; \gamma_i, \theta) / \partial \theta \). Let \( Z^\gamma_i \) denote the first derivative of \( Z \) with respect to \( \gamma_i^T \). Define \( W^\gamma_i \) and \( W^\theta \) similarly. Define

\[
H_{i,\gamma_i}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[Z^\gamma_i(w_{it}; \gamma_i(\theta), \theta)] \quad \text{and} \quad H_{i,\theta}(\theta) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[W^\gamma_i(\theta) + W^\gamma_i(\theta) \partial \gamma_i(\theta) / \partial \theta],
\]

where \( W^\gamma_i(\theta) = W^\gamma_i(w_{it}; \gamma_i(\theta), \theta) \) and \( W^\gamma_i(\theta) = W^\gamma_i(w_{it}; \gamma_i(\theta), \theta) \). For notational simplicity, let \( \max_i \) and \( \max_{i,t} \) abbreviate \( \max_{1 \leq i \leq N} \) and \( \max_{1 \leq i \leq N, 1 \leq t \leq T} \), 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 \( i, \{w_{it}, t \geq 1\} \) is stationary strong mixing with mixing coefficient \( \alpha_i(\cdot) \). Let \( \alpha(\cdot) \equiv \max_i \alpha_i(\cdot) \) satisfies \( \alpha(s) \leq c_\alpha \rho^s \) 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 \( \min_i \{\min_{\varsigma_i, ||\varsigma_i - \varsigma_i^0|| > \eta} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\phi(w_{it}; \varsigma_i) - \phi(w_{it}; \varsigma_i^0)]\} > \epsilon \) and \( \inf_{||\theta - \theta^0|| > \eta} \frac{1}{N} \sum_{i=1}^{N} [\Psi_i(\gamma_i(\theta), \theta) - \Psi_i(\gamma_i(\theta^0), \theta^0)] \) \( > \epsilon \), where \( \Psi_i(\gamma_i(\theta), \theta) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\phi(w_{it}; \gamma_i(\theta))]. \)

(iii) Let \( \Upsilon \) and \( \Theta \) denote the parameter space for \( \varsigma_i \) and \( \theta \), respectively. \( \Upsilon \) is compact and convex and the true value \( \varsigma_i^0 \) lies in the interior of \( \Upsilon \) for all \( i = 1, \ldots, N \).

(iv) For a \((p + r + q) \times 1 \) vector \( d = (d_1, \ldots, d_{p+r+q})^T \in \mathbb{N}^{p+r+q} \), we let \( |d| \) denote \( \sum_{j=1}^{p+r+q} d_j \). Let \( D^d \phi_{it}(w_{it}; \varsigma_i) = \partial^d \phi_{it}(w_{it}; \varsigma_i) / \partial \varsigma_{i_1} \partial \varsigma_{i_2} \ldots \partial \varsigma_{i_{p+r+q}}, \) where \( \varsigma_{i,j} \) denotes the \( j \)th element of \( \varsigma_i \). There is a non-negative real function \( M(\cdot) \) such that \( \sup_{\varsigma_i \in \Upsilon} \|D^d \phi_{it}(w_{it}; \varsigma_i)\| \leq M(w_{it}) \) and \( \|D^d \phi_{it}(w_{it}; \varsigma_i) - D^d \phi_{it}(w_{it}; \varsigma_i')\| \leq M(w_{it}) \|\varsigma_i - \varsigma_i'\| \) for all \( \varsigma_i, \varsigma_i' \in \Upsilon \) and \( |d| \leq 3 \), and \( \max_i \mathbb{E}[M(w_{it})] \leq c_M \) for some \( c_M < \infty \) and \( \kappa \geq 6 \).

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

(vi) \( NT^{1-K/2} \to c \in [0, \infty), (p/T)(\ln T)^{6} \to 0 \) and \( q = O(p) \) as \( (N, T) \to \infty \).

Assumptions A1(i)–(v) parallel Assumptions A1(i)–(v) in SSP. Assumption A1(i) imposes that \( w_{it} \)'s are independent across individuals and strong mixing over time. This condition is commonly
assumed in the nonlinear panel literature; see, e.g., Hahn and Kuersteiner (2011) and SSP. The stationarity condition is not necessary; it is assumed only for the purpose of simplifying the notation in the proofs of some asymptotic results in the appendix. Note that Assumption A1(i) allows the dynamic panel data model. Assumption A1(ii) imposes the identification condition for the common parameter \( \theta \). Assumption A1(iii) requires \( \{c_i\} \) 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 \( H_{i,\gamma}(\theta) \) and \( H_{i,\theta_0}(\theta^0) \) 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 \( \theta^0 \) 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) \( \| \tilde{\theta} - \theta^0 \| = O_P \left( (q(T)^{1/2} \right) , (ii) \| \tilde{\gamma}_i - \gamma_i^0 \| = O_P \left( (p(T)^{1/2} \right) , (iii) \max_{1 \leq i \leq N} \| \tilde{\gamma}_i - \gamma_i^0 \| = O_P((p(T)^{1/2} (\ln T)^{3}) \), 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, Shi, and Phillips (2016b, SSPb hereafter) we can establish that \( P(\max_{1 \leq i \leq N} \| \tilde{\gamma}_i - \gamma_i^0 \| \geq C(p/T)^{1/2} (\ln T)^3 \) = o\((N^{-1})\) for some large positive constant \( C \). We can obtain a slightly tighter probability order for \( \max_{1 \leq i \leq N} \| \tilde{\gamma}_i - \gamma_i^0 \| \) 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 \( \theta \) 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 is 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 \( |G_k(K)| \geq 1 \) for \( k = 1, \ldots, K \). Define \( \hat{\sigma}_{\mathcal{G}(K)}^2 = 2(NT)^{-1} \sum_{k=1}^{K} \sum_{i \in G_k} \sum_{t=1}^{T} \varphi(w_{it}; \hat{\beta}_i(K), \hat{\mu}_i(K), \hat{\theta}(K)) \), where \( \hat{\beta}_i(K), \hat{\mu}_i(K) \), and \( \hat{\theta}(K) \) solve the constrained problem in (2.7) with \( \{G_k(K)\} \) being replaced by \( \{G_k(K)\} \).

We add two assumptions.

**Assumption A2** (i) There exists a constant \( c_L > 0 \) such that \( \min_{1 \leq k < k' \leq K^0} \| \alpha_k^0 - \alpha_{k'}^0 \| > c_L \).
(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) $p^{3/2}N \sqrt{\log N}/N \to 0$ as $(N, T) \to \infty$.

(ii) As $(N, T) \to \infty$, let $\min_{1 \leq k \leq K^0} \min_{g(K)} \sigma^2_{\hat{g}(K)} \to \sigma^2$, where $\sigma^2 = \lim_{(N, T) \to \infty} 2(NT)^{-1} \sum_{k=1}^{K^0} \sum_{i \in G_k^0} \sum_{t=1}^{T} \mathbb{E} \varphi(w_{it}; \alpha^0_k, \theta^0)$. 

(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 is bounded away from zero. At the cost of more complicated arguments, we can allow $\min_{1 \leq k \leq K^0} \| \alpha^0_k - \alpha^0_k' \|$ 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 below while KLZ only apply a rough probability bound to control them. Assumption A3(ii)–(iii) imposes some typical conditions to ensure both over-grouped and under-grouped panel structure models are ruled out. In particular, Assumption A3(ii) ensures that for all under-fitted models, the mean square errors would be asymptotically greater than $\sigma^2_0$.

The following theorem indicates that we can estimate the true group structure $G^0$ 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 $\hat{G}(K^0) = \{ \hat{G}_1(K^0), \ldots, \hat{G}_{K^0}(K^0) \}$ be the estimated group structure based on the SBSA 1. Then $P(\hat{G}(K^0) = G^0) \to 1$ 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 $G^0$ 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 $G^0$ 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 $\rho_{NT}$ can only converge to zero at a speed slower than $T^{-1}$. This is simply due to the fact that the heterogeneous incidental parameters $\mu_i$’s in the model can only be estimated at the slow $T^{-1/2}$ convergence rate. For linear panel data models where $\mu_i$ 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*)** $pp_{NT} \to 0$ as $(N,T) \to \infty$ and $(NT + T^2)\rho_{NT} \to \infty$ as $(N,T) \to \infty$.

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

**Assumption A3 (iii**) $pp_{NT} \to 0$ 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 $U_{it}(w_{it}; \beta_i, \mu_i, \theta) \equiv \partial \varphi(w_{it}; \beta_i, \mu_i, \theta)/\partial \beta_i$ and $V(w_{it}; \beta_i, \mu_i, \theta) \equiv \partial \varphi(w_{it}; \beta_i, \mu_i, \theta)/\partial \mu_i$. Let $U_j$ denotes the jth element in $U$, and similarly for $V_j$ and $W_j$. Let $U^\beta$ denote the derivative of $U$ with respect to $\beta^T$. Define $U^\mu_i, V_i, V^\theta, W_i$ and $W^\theta_i$ analogously. For notational simplicity, let $U_{it} \equiv U(w_{it}; \beta_i^0, \mu_i^0, \theta^0)$, and similarly for $V_{it}, W_{it}$, $U_{it}^\mu, V_{it}, V_{it}^\theta, W_{it}^\mu$ and $W_{it}^\theta$. Let $U_{it,j}^\mu \equiv \partial U_j(w_{it}; \beta_i^0, \mu_i^0, \theta^0)/\partial \mu_{it,j}^T$ and $U_{it,j}^{\mu T} \equiv \partial^2 U_j(w_{it}; \beta_i^0, \mu_i^0, \theta^0)/\partial \mu_{it,j} \partial \mu_{it,j}^T$, and similarly for $W_{it,j}^\mu, V_{it,j}^\mu$ and $W_{it,j}^{\mu T}$. Define $S_{it}^U = \frac{1}{T} \sum_{t=1}^T E(U_{it}^\mu), S_{it}^V = \frac{1}{T} \sum_{t=1}^T E(V_{it}^\mu)$, $S_{it}^W = \frac{1}{T} \sum_{t=1}^T E(W_{it}^\mu)$, $S_{it,2, j}^{U} = \frac{1}{T} \sum_{t=1}^T E(U_{it,j}^\mu)$, $S_{it,2, j}^{V} = \frac{1}{T} \sum_{t=1}^T E(V_{it,j}^\mu), S_{it,2, j}^{W} = \frac{1}{T} \sum_{t=1}^T E(W_{it,j}^\mu)$, $U_{it} = U_{it}^\mu - S_{it}^U S_{it}^{-1} V_{it}, V_{it}^\mu = U_{it}^\mu - S_{it}^U S_{it}^{-1} V_{it}, W_{it} = U_{it}^\mu - S_{it}^U S_{it}^{-1} V_{it}, W_{it}^\mu = U_{it}^\mu - S_{it}^U S_{it}^{-1} V_{it}, \Omega_{U,T,\beta} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^T E(U_{it} W_{it}^T), \Omega_{U,T,\theta} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^T E(U_{it} W_{it}^T),$ and $\Omega_{U,T,\theta} = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^T E(U_{it} W_{it}^T)$.

Next, let $B_{NT} = (B_{1, NT}, \ldots, B_{K, NT}, B_{\theta, NT})^T = ((B_{1, NT} - B_{2, NT})^T, \ldots, (B_{1, K, NT} - B_{2, K, NT})^T)^T, \ldots, (B_{1, K, NT} - B_{2, K, NT})^T)^T, (B_{1, \theta, NT} - B_{2, \theta, NT})^T)$, where

$$B_{1, NT} = \frac{1}{\sqrt{N_k T^2}} \sum_{i \in G_k} \sum_{s=1}^T \sum_{t=1}^T U_{it}^\mu S_{it}^{-1} V_{is}, B_{1, \theta, NT} = (N T^3)^{-1/2} \sum_{i=1}^N \sum_{s=1}^T \sum_{t=1}^T \sum_{i=1}^T S_{it}^U S_{it}^{-1} V_{is},$$

$$[B_{2, NT}]_j = \frac{1}{2 \sqrt{N_k T}} \sum_{i \in G_k} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it} \right)^T S_{it}^{-1} S_{it}^{U,2, j} S_{it}^{-1,1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it} \right) - \frac{1}{2 \sqrt{N_k T}} \sum_{i \in G_k} S_{it}^U S_{it}^{-1} R_{it},$$

$$[B_{2, \theta, NT}]_j = \frac{1}{2 \sqrt{N T}} \sum_{i=1}^N \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it} \right)^T S_{it}^{-1} S_{it}^{W,2, j} S_{it}^{-1,1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it} \right) - \frac{1}{2 \sqrt{N T}} \sum_{i=1}^N S_{it}^W S_{it}^{-1} R_{it},$$

15
where \([A]_j\) denotes the \(j\)th 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 \([R_{iW}]_j = (\frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it})^\top S_{IV}^{-1} S_{IV2,j} S_{IV}^{-1} (\frac{1}{\sqrt{T}} \sum_{t=1}^T V_{it})\). Define

\[
\Omega_{NT} = \begin{bmatrix}
\frac{1}{N_1} \sum_{i \in G_1^0} \Omega_{iT, \beta \beta} & \cdots & 0 & \frac{1}{N_1} \sum_{i \in G_1^0} \Omega_{iT, \beta \theta} \\
\vdots & \ddots & \vdots & \vdots \\
0 & \cdots & \frac{1}{N_{K_0}} \sum_{i \in G_{K_0}^0} \Omega_{iT, \beta \beta} & \frac{1}{N_{K_0}} \sum_{i \in G_{K_0}^0} \Omega_{iT, \beta \theta} \\
\frac{1}{N_1} \sum_{i \in G_1^0} \Omega_{iT, \beta \theta} & \cdots & \frac{1}{N_{K_0}} \sum_{i \in G_{K_0}^0} \Omega_{iT, \beta \theta} & \frac{1}{N} \sum_{i=1}^N \Omega_{iT, \theta \theta}
\end{bmatrix}
\]

and

\[
\mathbb{H}_{NT} (\beta, \theta) = \begin{bmatrix}
\frac{1}{N_1} \sum_{i \in G_1^0} H_{i, \beta \beta} (\beta, \theta) & \cdots & 0 & \frac{1}{N_1} \sum_{i \in G_1^0} H_{i, \beta \theta} (\beta, \theta) \\
\vdots & \ddots & \vdots & \vdots \\
0 & \cdots & \frac{1}{N_{K_0}} \sum_{i \in G_{K_0}^0} H_{i, \beta \beta} (\beta, \theta) & \frac{1}{N_{K_0}} \sum_{i \in G_{K_0}^0} H_{i, \beta \theta} (\beta, \theta) \\
\frac{1}{N} \sum_{i=1}^N H_{i, \theta \beta} (\beta, \theta) & \cdots & \frac{1}{N} \sum_{i=1}^N H_{i, \theta \theta} (\beta, \theta)
\end{bmatrix},
\]

where

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

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

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

and

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

Let \(\mathbb{H}_{NT} \equiv \mathbb{H}_{NT} (\beta^0, \theta^0)\). 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 \(\alpha_k\) and \(\theta\) that are obtained with \(\hat{K}\) and \(\hat{G}_k (\hat{K})\) in (2.10) being replaced by \(K^0\) and \(G_0^k\). The following theorem indicates that these two sets 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

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

where $D_{NT} = \text{diag}(\sqrt{N_1T_{I_p}}, \ldots, \sqrt{N_{K^0}T_{I_p}}, \sqrt{N_1T_{I_q}})$, $S$ is an arbitrary $k \times (pK^0 + q)$ 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 $S$ 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 $S$ 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 formulae are a little bit more complicated than theirs due to the presence of the common parameter $\theta$. In the absence of $\theta$, both formulae can be simplified and one can easily verify that in this case the asymptotic bias and variance of $\hat{\alpha}_k$’s are the same as those of the group-specific parameter estimators in SSP. Following Hahn and Newey (2004), 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 formulae for both estimates and justify their consistency. Alternatively, we can use the jackknife method to correct bias. See Hahn and Newey (2004) and Dhaene and Jochmans (2015) for static and dynamic panels, respectively.

### 4 An improved algorithm

In this section, we consider an improved algorithm that is based on the spectral decomposition of the $N \times N$ matrix $\tilde{D}_N = N^{-1}\tilde{\beta}\tilde{\beta}^\top$. We first explain why the eigenvectors associated with the few largest eigenvalues of $\tilde{D}_N$ 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:

\[ A = \alpha^0 \alpha^{0\top} \quad \text{and} \quad D_N = N^{-1} \beta^0 \beta^{0\top}. \]  

Define an $N \times K^0$ matrix $Z_N \in \{0,1\}^{N \times K^0}$ 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 $i$th row of $Z_N$ 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

\[ D_N = N^{-1} Z_N A Z_N^\top. \]  

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, $Z_N$ stores the individuals’ group membership in our model and nodes’ community membership in an SBM. The matrix $A$ 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 $A$ to lie between 0 and 1. In both cases, the main interest is to estimate $Z_N$ 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, Chatterjee, and Yu (2011) work on the eigenvectors of a normalized adjacency matrix. This motivates us to consider the eigenvectors of the sample analogue of $D_N$, 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 $\tilde{\beta}$. If the difference between different columns of the $p \times K^0$ matrix $\alpha^{0\top}$ 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 $N^{-1} \tilde{\beta} \tilde{\beta}^\top$ (or $D_N$) summarize all the useful group information and implementing the SBSA on the eigenvectors tend to outperform that based on the original $\tilde{\beta}$ matrix. Due to limited space, we only consider $N = 200$ and $T = 20$ for a linear DGP with three groups ($K^0 = 3$) and $p$ regressors, where the group ratio is $3 : 3 : 4$. We consider three values of $p$: 6, 8, 10. In Figure 1, the first row plots different columns in $\tilde{\beta}$ for
Figure 1: Comparison of the plots of the $p$ columns in the preliminary estimates $\tilde{\beta}$ with the three eigenvectors of $N^{-1}\tilde{\beta}\tilde{\beta}^T$ associated with its three largest eigenvalues when $p > K^0$: row 1 for preliminary estimates and row 2 for eigenvectors

$p = 6, 8, 10$, and the second row plots the three eigenvectors corresponding to the three largest eigenvalues of $N^{-1}\tilde{\beta}\tilde{\beta}^T$ 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 $\tilde{\beta}$. 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 $\alpha^{0\top}$ are linearly dependent. Let $p = 2$ and $K^0 = 3$. The true group-specific parameters are

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

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 $N^{-1}\tilde{\beta}\tilde{\beta}^T$ contains all the useful individual group identity information in $\tilde{\beta}$ and essentially serves as a “signal enhancement”. As a result, such an eigenvector can reveal the true group information much more clearly than $\tilde{\beta}$ itself.

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

$A = u\Lambda u^\top$,

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

\[
D_N = N^{-1} U_N \Sigma_N U_N^\top = N^{-1} U_{1,N} \Sigma_1,N U_{1,N}^\top,
\]

where \( \Sigma_N = \text{diag}(\mu_1, \ldots, \mu_{K^*}, 0, \ldots, 0) \) is a \( p \times p \) matrix that contains the eigenvalues of \( D_N \) in descending order along its diagonal, \( \Sigma_1,N = \text{diag}(\mu_1, \ldots, \mu_{K^*}) \), the columns of \( U_N \) contain the eigenvectors of \( D_N \) associated with the eigenvalues in \( \Sigma_N \), \( U_N = (U_{1,N}, U_{2,N}) \), and \( U_N^\top U_N = I_p \).

The following lemma establishes the link between the eigenvalues and eigenvectors of \( A \) and those of \( D_N \).

**Lemma 4.1.** Let \( A, D_N, \Lambda, \Sigma_1,N, u \) and \( U_{1,N} \) be defined as above. Then there exists a nonsingular matrix \( S \equiv S_N \) such that (i) the diagonal matrix \( \Sigma_1,N \) can be written as \( S^{-1} \Lambda (S^{-1})^\top \), (ii) \( U_{1,N} = N^{-1/2} Z_N u S \), (iii) \( S \) is given by \((N^{-1/2} U_{1,N}^\top Z_N u)^{-1} \), and (iv) \( z_i^\top u S = z_j^\top u S \) if and only \( z_i = z_j \) for \( i, j = 1, 2, \ldots, N \).

The last result in Lemma 4.1 is obvious if \( u S \) is a nonsingular square matrix. In this case, there exists a one-to-one map between \( U_{1,N} \) and \( Z_N \). In the general case, we allow \( K^* < K^0 \) so that \( u S \) has rank \( K^* \) only, and we show in the proof of the above lemma that the rows of \( u S \) are distinct from each other. This ensures that the rows of \( U_{1,N} \) contain the same group information as \( Z_N \).

Therefore, we can infer each individual’s group membership based on the eigenvector matrix \( U_{1,N} \) if \( D_N \) is observed.

In practice, \( D_N \) is not observed. But we can estimate it by

\[
\tilde{D}_N \equiv N^{-1} \hat{\beta} \hat{\beta}^\top.
\]

Consider the spectral decomposition of \( \tilde{D}_N: \tilde{D}_N = \tilde{U}_N \tilde{\Sigma}_N \tilde{U}_N^\top \), where \( \tilde{\Sigma}_N = \text{diag}(\tilde{\mu}_1, \ldots, \tilde{\mu}_p,N) \) contains the first \( p \) eigenvalues of \( \tilde{D}_N \) in descending order. By Theorem 3.1, we can readily show that \( \| \tilde{D}_N - D_N \| = O_P ((p/T)^{1/2}) \), ensuring that \( \max_{1 \leq \ell \leq N} |\tilde{\mu}_{\ell,N} - \mu_{\ell,N}| \leq \| \tilde{D}_N - D_N \| = O_P ((p/T)^{1/2}) \), where \( \tilde{\mu}_{\ell,N} \) and \( \mu_{\ell,N} \) denote the \( \ell \)th largest eigenvalues of \( \tilde{D}_N \) and \( D_N \), respectively.

To take into account the possibility of estimating a zero eigenvalue of \( D_N \) by a positive value, we have to ensure that \( \mu_{K^*} \) is not too close to zero in order to identify the nonzero eigenvalues of \( D_N \) 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, Chatterjee, and Yu (2011), Theorem 3 in Yu, Wang, and Samworth (2015)).

Recall \( \lambda_j(A) \) denotes the \( j \)th 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 \( \lambda_{K^*}(A) = \lambda_{K^*}(a^0 a^{0\top}) \geq c \) for some constant \( c > 0 \). Noting that \( AB \) and \( BA \) share the same set of nonzero eigenvalues, we have

\[
\begin{align*}
\mu_{K^*} = & \lambda_{K^*}(D_N) = \lambda_{K^*}(N^{-1} Z_N A Z_N^\top) = \lambda_{K^*}(A N^{-1} Z_N^\top Z_N) \\
\geq & \lambda_{K^*}(A) \lambda_{\min} \left(N^{-1} Z_N^\top Z_N\right) \geq c \min_{1 \leq k \leq K^0} N_k / N. 
\end{align*}
\]  
(4.3)
It follows that \( \lim_{N \to \infty} \mu_{K^*N} \geq \epsilon \min_{1 \leq k \leq K^0} \tau_k > 0 \) under Assumption A2(ii). Since only the eigenvectors that are associated with the \( K^* \) nonzero eigenvalues of \( \tilde{D}_N \) can contain the group information, we will restrict our attention to the eigenvectors associated with the first \( \kappa_N \) eigenvalues of \( \tilde{D}_N \) such that \( \lambda_{\kappa_N}(\tilde{D}_N) \geq c_N \), 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 \( \tilde{D}_N \) whose population values are zero. To see this, notice that when \( \kappa_N > K^* \), \( \lambda_{\kappa_N}(\tilde{D}_N) \) converges to zero in probability at rate \( (p/T)^{1/2} \). So it is easy to show that \( \kappa_N = K^* \) w.p.a.1.

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

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

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

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

**Theorem 4.2.** Suppose that Assumptions A1–A5 hold and \( \max_{1 \leq i \leq N} \| \beta_i^0 \| = O(1) \). Then \( \kappa_N = K^* \) w.p.a.1. Furthermore, conditional on \( \kappa_N = K^* \), there exists a sequence of \( K^* \times K^* \) orthogonal matrices \( O_N \) such that \( \max_{1 \leq i \leq N} \sqrt{N} \| \tilde{u}_{1,i} - O_N u_{1,i} \| = O_P \left( (p/T)^{1/2} \ln T \right) \).

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

### 4.2 An eigenvector-based SBSA

Since \( \tilde{U}_{1,N} \) contains the group membership for all individuals, we implement the SBSA based on it. Let \( \tilde{U}_{1,N} = (\tilde{U}_{1,1}, \ldots, \tilde{U}_{1,\kappa_N}) \) and \( U_{1,N} = (U_{1,1}, \ldots, U_{1,\kappa_N}) \). Let \( U_{ij} \) and \( \tilde{U}_{ij} \) denote the \( i \)th element of \( U_j \) and \( \tilde{U}_j \), respectively. We sort the \( N \) elements of \( \tilde{U}_j \) in ascending order and denote the order statistics by

\[
\tilde{U}_{\pi_j(1),j} \leq \tilde{U}_{\pi_j(2),j} \leq \cdots \leq \tilde{U}_{\pi_j(N),j},
\]

(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

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

where \( 1 \leq i < l \leq N \).

Let

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

To account for the scale effect, we use \( \tilde{\beta}' = (\tilde{\beta}'_1, \ldots, \tilde{\beta}'_p) \) where \( \tilde{\beta}'_j = \tilde{\beta}_j / \sqrt{\sigma_{\pi_j(j)}^2} \), \( j = 1, \ldots, p \), instead of \( \tilde{\beta} \) in calculating the eigenvectors \( \tilde{U}_{1,N} \). Recall that \( \sigma_{\pi_j(j)}^2 \) is defined in Section 2.2.
denote the sample mean and variance of the subsample $\tilde{S}_{i,l}(j)$. Define

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

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

**Sequential Binary Segmentation Algorithm 2 (SBSA 2).** SBSA 2 is essentially the same as with $\tilde{U}_{1,N} (N \times K_N)$, $\tilde{S}_{i,l}(j)$, $\tilde{U}_{i,l}(j)$, $\tilde{V}_{i,l}(j)$, and $\tilde{S}_{i,l}(j,m)$ in place of $\tilde{\beta} (N \times p)$, $S_{i,l}(j)$, $\beta_{i,l}(j)$, $V_{i,l}(j)$, and $\tilde{S}_{i,l}(j,m)$ 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 $\hat{G} (K^0) \equiv \{ \hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{K^0} \}$ as the estimates of the true group structure $G^0$. Otherwise, we can estimate $K^0$ either based on SBSA 1 or SBSA 2.

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

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

Let

$$\hat{K} \equiv \arg \min_{1 \leq K \leq K^{\max}} \text{IC}_2(K) \text{ and } \hat{G} \equiv \hat{G}(\hat{K}) \equiv \{ \hat{G}_1(\hat{K}), \hat{G}_2(\hat{K}), \ldots, \hat{G}_{\hat{K}}(\hat{K}) \}, \quad (4.7)$$

be the estimated number of groups and the estimated group structure, respectively. We will show that $P(\hat{K} = K^0) \rightarrow 1$ and $P(\hat{G} = G^0) \rightarrow 1$ as $(N,T) \rightarrow \infty$.

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

### 4.3 Asymptotic properties

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

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

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

\(^4\)For completeness, we provide the details of SBSA 2 in the online supplementary material.
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 $\hat{\alpha}_k$’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 $\{\hat{K} = K_0\}$, the asymptotic distribution of $D_N(\hat{\alpha}_1 - \alpha_1^0)^T, \ldots, (\hat{\alpha}_K - \alpha_K^0)^T, (\hat{\theta} - \theta^0)^T$ is identical to $D_N((\alpha_1 - \alpha_1^0)^T, \ldots, (\alpha_K - \alpha_K^0)^T, (\theta - \theta^0)^T)$ studied in Theorem 3.4.

Combining the results in Theorems 4.3–4.4, we can recover the true group structure $G^0$ w.p.a.1 by using the SBSA 2 and IC2 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 $\hat{\beta}_i$ and the convergence rate does not play a role here. Theorem 4.2 ensures that $\hat{U}_{1,N}$ contains all the individuals’ group information that is required and it implies the uniform convergence of $\sqrt{N}(\hat{u}_{1,i} - Ou_{1,i})$ 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 $\hat{G} = \hat{G}(\hat{K})$ with $G^0$, 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

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

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

$$(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \begin{pmatrix} 0.5 \\ -1 \\ 1 \\ 0.5 \\ 2 \end{pmatrix}.$$
Note that here \( \alpha_{0,1}^0 = \alpha_{2,1}^0 = \alpha_{3,1}^0 \) 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 \( \times \) 1 vector with the \( j \)th element given by \( x_{j, it} = 0.2 \mu_i + \varepsilon_{j, it}, j = 1, \ldots, 10 \), and \( \varepsilon_{j, it}, \varepsilon_{it} \), and the fixed effect \( \mu_i \) are all i.i.d. standard normal and mutually independent of each other. The true coefficients \( \beta_i \) 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 \( j \)th element given by \( x_{j, it} = 0.2 \mu_i + \varepsilon_{j, it}, j = 1, \ldots, p \), and \( \varepsilon_{j, it}, \varepsilon_{it}, \) and the fixed effect \( \mu_i \) 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 \( \beta_i \) can be classified into 3 groups with true group-specific parameter values given by

\[
(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left( \begin{array}{ccc}
-1 & 0.5 & 1 \\
-0.9 & 0.6 & 1.1 \\
\vdots & \vdots & \vdots \\
-1 + 0.1(p - 1) & 0.5 + 0.1(p - 1) & 1 + 0.1(p - 1)
\end{array} \right).
\]

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

\[
(\alpha_1^0, \alpha_2^0, \alpha_3^0) = \left( \begin{array}{ccc}
1.5 & -0.5 & -1.8 \\
-1.5 & 0.5 & 1.8
\end{array} \right).
\]

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

\[
y_{it} = \max \left( 0, p y_{it-1} + x_{it}^\top \beta_i + \mu_i + \varepsilon_{it} \right),
\]

24
where \( x_{it}, \mu_i, \) and \( \varepsilon_{it} \) 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

\[
(\alpha^0_1, \alpha^0_2, \alpha^0_3) = \begin{pmatrix}
-1.2 \\
1.6 \\
0.6 \\
-0.8 \\
1.5 \\
-1.9
\end{pmatrix}.
\]

As in DGP 4, \( \sigma^2 \) is modeled as the common parameter across all individuals but we do not assume \( \rho \) 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

\[
IC_1(K) = \sigma^2_{\hat{G}(K)} + pK\rho_1(NT),
\]

where \( \rho_1(NT) = \frac{1}{30} \ln(NT)/(NT)^{1/3} \), \( \hat{G}(K) = \{ \hat{G}_1(K), \ldots, \hat{G}_K(K) \} \), \( \sigma^2_{\hat{G}(K)} = \frac{1}{NT} \sum_{k=1}^{K} \sum_{i \in \hat{G}_k(K)} \sum_{t=1}^{T} (\tilde{y}_{it} - \tilde{x}_{it}^T \alpha_k(K))^2 \), \( \tilde{y}_{it} = y_{it} - T^{-1} \sum_{t=1}^{T} y_{it} \), and similarly for \( \tilde{x}_{it} \). For DGPs 4–5, the information criterion is

\[
IC_2(K) = 2L_{NT}(\hat{\beta}(K), \hat{\mu}(K), \hat{\theta}(K)) + pK\rho_2(NT),
\]

where \( L_{NT}(\cdot) \) is given in Section 2, and \( \rho_2(NT) = \frac{1}{30} \ln(NT)/(NT)^{1/3} \).

### 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 \( K^{\text{max}} = 5 \). The true number of groups is given by \( K^0 = 3 \). We compare 4 algorithms: K-means on \( \hat{\beta} \), K-means on the eigenvectors of \( N^{-1}\tilde{\beta}\tilde{\beta}^T \), 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, 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.

\(^5\)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 $K = 1, \ldots, 5$ groups when $K^0 = 3$ and $K^{\text{max}} = 5$

<table>
<thead>
<tr>
<th></th>
<th>DGP 1</th>
<th></th>
<th>DGP 2</th>
<th></th>
<th>DGP 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 2 3</td>
<td>1 2 3</td>
<td>1 2 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$N$</td>
<td>$T$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0 0</td>
<td>0.880 0.120 0</td>
<td>0.588 0.372 0.040 0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0 0</td>
<td>0.874 0.126 0</td>
<td>0.102 0.876 0.022 0</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0 0</td>
<td>0.906 0.094 0</td>
<td>0.090 0.838 0.072 0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0 0</td>
<td>0.872 0.116 0.012 0</td>
<td>0.278 0.678 0.044 0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0 0</td>
<td>0.854 0.146 0</td>
<td>0.228 0.694 0.078 0</td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0 0</td>
<td>0.870 0.128 0.002 0</td>
<td>0.198 0.722 0.080 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>0 0</td>
<td>0.284 0.200 0.210 0.306</td>
<td>0.532 0.186 0.166 0.116</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0 0</td>
<td>0.060 0.414 0.348 0.178</td>
<td>0.212 0.384 0.170 0.234</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0 0</td>
<td>0.794 0.198 0.008</td>
<td>0.090 0.156 0.354 0.400</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0 0</td>
<td>0.146 0.234 0.274 0.346</td>
<td>0.700 0.160 0.118 0.022</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0 0</td>
<td>0.022 0.342 0.378 0.258</td>
<td>0.492 0.354 0.106 0.048</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0 0</td>
<td>0.734 0.262 0.004</td>
<td>0.224 0.290 0.264 0.222</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.012 0.982 0.006 0</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.012 0.982 0.006 0</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.012 0.982 0.006 0</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.012 0.982 0.006 0</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.012 0.982 0.006 0</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.012 0.982 0.006 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>0 0</td>
<td>0.012 0.988 0</td>
<td>0.282 0.274 0.254 0.190</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0 0</td>
<td>0.012 0.988 0</td>
<td>0.282 0.274 0.254 0.190</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0 0</td>
<td>0.012 0.988 0</td>
<td>0.282 0.274 0.254 0.190</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0 0</td>
<td>0.012 0.988 0</td>
<td>0.282 0.274 0.254 0.190</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0 0</td>
<td>0.012 0.988 0</td>
<td>0.282 0.274 0.254 0.190</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0 0</td>
<td>0.012 0.988 0</td>
<td>0.282 0.274 0.254 0.190</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.088 0.898 0.014 0</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.088 0.898 0.014 0</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.088 0.898 0.014 0</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.088 0.898 0.014 0</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.088 0.898 0.014 0</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0 0</td>
<td>0.996 0.004 0</td>
<td>0.088 0.898 0.014 0</td>
</tr>
</tbody>
</table>
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 $N^{-1} \sum_{k=1}^{K^0} \sum_{i \in G_k} 1\{\beta_i^0 = \alpha_{k,i}^0\}$, 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 $\alpha_{k,2}^0 \equiv (\alpha_{k,1,2}^0, \ldots, \alpha_{K^0,2}^0)^T$, 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 $\alpha_{k,2}^0$; $k = 1, \ldots, K^0$, with weight $N_k/N$. Specifically, it is $\sum_{k=1}^{K^0} \frac{N_k}{N} \text{RMSE}(\alpha_{k,2}^0)$. Similarly, we define weighted versions of bias, and coverage ratio of the 95% confidence interval estimators. Tables 3 and 4 contain the classification and post-classification results where the oracle estimates are obtained by using the true group structure and the other estimates are obtained based 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.
Table 3: Classification and point estimation of $\alpha_2^0$

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th>Correct Comparison Criteria</th>
<th>Correct Comparison Criteria</th>
<th>Correct Comparison Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oracle</td>
<td>K-means</td>
<td>SBSA 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct Comparison Criteria</td>
<td>Correct Comparison Criteria</td>
<td>Correct Comparison Criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oracle</td>
<td>K-means</td>
<td>SBSA 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct Comparison Criteria</td>
<td>Correct Comparison Criteria</td>
<td>Correct Comparison Criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C-Lasso</td>
<td>SBSA 1</td>
<td>SBSA 2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correct Comparison Criteria</td>
<td>Correct Comparison Criteria</td>
<td>Correct Comparison Criteria</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N</td>
<td>T</td>
<td>N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DGP 1</th>
<th>DGP 2</th>
<th>DGP 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Oracle</th>
<th>100</th>
<th>1</th>
<th>0.061</th>
<th>0.002</th>
<th>0.924</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1</td>
<td>0.040</td>
<td>0.002</td>
<td>0.917</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>0.027</td>
<td>0.000</td>
<td>0.958</td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>0.041</td>
<td>0.001</td>
<td>0.931</td>
<td></td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1</td>
<td>0.027</td>
<td>0.000</td>
<td>0.950</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>1</td>
<td>0.019</td>
<td>0.001</td>
<td>0.946</td>
</tr>
<tr>
<td>K-means</td>
<td>100</td>
<td>0.884</td>
<td>0.301</td>
<td>-0.081</td>
<td>0.764</td>
</tr>
<tr>
<td>on  $\beta$</td>
<td>100</td>
<td>0.914</td>
<td>0.293</td>
<td>-0.106</td>
<td>0.787</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.960</td>
<td>0.214</td>
<td>-0.055</td>
<td>0.886</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.894</td>
<td>0.247</td>
<td>-0.072</td>
<td>0.760</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.923</td>
<td>0.272</td>
<td>-0.091</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.945</td>
<td>0.227</td>
<td>-0.067</td>
<td>0.854</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.702</td>
<td>0.537</td>
<td>-0.243</td>
<td>0.334</td>
</tr>
<tr>
<td>on eigen- $\beta$</td>
<td>100</td>
<td>0.778</td>
<td>0.383</td>
<td>-0.191</td>
<td>0.491</td>
</tr>
<tr>
<td>vectors</td>
<td>100</td>
<td>0.902</td>
<td>0.278</td>
<td>-0.106</td>
<td>0.794</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.695</td>
<td>0.544</td>
<td>-0.254</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.775</td>
<td>0.376</td>
<td>-0.171</td>
<td>0.430</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.873</td>
<td>0.316</td>
<td>-0.142</td>
<td>0.735</td>
</tr>
<tr>
<td>C-Lasso</td>
<td>100</td>
<td>0.941</td>
<td>0.076</td>
<td>-0.015</td>
<td>0.861</td>
</tr>
<tr>
<td>SBSA 1</td>
<td>100</td>
<td>0.986</td>
<td>0.044</td>
<td>-0.006</td>
<td>0.900</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.999</td>
<td>0.027</td>
<td>-0.000</td>
<td>0.957</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.942</td>
<td>0.053</td>
<td>-0.018</td>
<td>0.829</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.986</td>
<td>0.028</td>
<td>-0.005</td>
<td>0.938</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.999</td>
<td>0.019</td>
<td>0.001</td>
<td>0.937</td>
</tr>
<tr>
<td>SBSA 2</td>
<td>100</td>
<td>0.930</td>
<td>0.089</td>
<td>0.006</td>
<td>0.851</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.984</td>
<td>0.043</td>
<td>-0.003</td>
<td>0.901</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.999</td>
<td>0.027</td>
<td>0.000</td>
<td>0.959</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.932</td>
<td>0.052</td>
<td>0.003</td>
<td>0.856</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.985</td>
<td>0.029</td>
<td>-0.001</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.999</td>
<td>0.019</td>
<td>0.001</td>
<td>0.943</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.931</td>
<td>0.077</td>
<td>0.005</td>
<td>0.860</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.985</td>
<td>0.043</td>
<td>-0.003</td>
<td>0.911</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.998</td>
<td>0.027</td>
<td>-0.000</td>
<td>0.958</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.930</td>
<td>0.051</td>
<td>0.005</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.984</td>
<td>0.030</td>
<td>-0.002</td>
<td>0.925</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.999</td>
<td>0.019</td>
<td>0.001</td>
<td>0.942</td>
</tr>
</tbody>
</table>
Table 4: Classification and point estimation of $\alpha_2^0$

<table>
<thead>
<tr>
<th>DGP 4</th>
<th>DGP 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>(T)</td>
</tr>
<tr>
<td>()</td>
<td>()</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Oracle</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>K-means on $\tilde{\beta}$</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>K-means on eigen-vectors</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>SBSA 1</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td>SBSA 2</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>200</td>
</tr>
</tbody>
</table>
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 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 $\tilde{\beta}$ 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, Gomes, and Maenhout (2005) investigate how labor income and financial wealth affect portfolio decisions. Unobservable factors also play a very important role in the process of portfolio decision making. For example, individual risk preference, habits and information acquirement affect how people respond to various observable factors. Samuelson (1969) models risk preference as the fundamental factor in portfolio choices. Polkovnichenko (2007) employs 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,
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

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

where $x_{1,it}$ includes log financial assets and log non-capital income, $x_{2,it}$ includes AEX premium, time trend and retirement dummy, $\mu_i$ is the fixed effect, and $\varepsilon_{it}$’s are i.i.d. normal. The observable dependent variable $y_{it}$ is subject to two-sided censoring: $y_{it} = \max(0, y_{it}^*, 1)$. Note that $\beta_2$ is common across individuals in (6.1). We assume that the true values of $\beta_{1i}$’s exhibit the group structure, $\beta_{1i}^0 = \sum_{k=1}^{K_0} \alpha_k^0 \cdot 1 \{ i \in G_k^0 \}$. 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 $\beta_{1i}$’s to be heterogeneous across groups and impose homogeneity assumption on $\beta_2$. 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 $i$. 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

---

6AEX 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, Hochguertel, and Van Soest (2002) and AS.

7Note 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.
Figure 2: Scatter plot for the preliminary estimates of $\beta_i$’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.

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.

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

### 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. Figure 2 reports the scatter diagram for
Table 6: Regression results for the DNB household survey dataset

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

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. $^a$, $^b$, and $^c$ denote significance at 10%, 5% and 1% levels, respectively.

The preliminary estimates of $\beta_{1i}$, 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. 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 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 $T_i$.

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

<table>
<thead>
<tr>
<th></th>
<th>$T_i \geq 9$</th>
<th></th>
<th>$T_i \geq 8$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1</td>
<td>Group 2</td>
<td>Group 3</td>
<td>Group 1</td>
</tr>
<tr>
<td>log(FA)</td>
<td>-0.043$^c$</td>
<td>-0.240$^c$</td>
<td>-0.055$^c$</td>
<td>-0.040$^c$</td>
</tr>
<tr>
<td></td>
<td>(0.008)</td>
<td>(0.009)</td>
<td>(0.007)</td>
<td>(0.008)</td>
</tr>
<tr>
<td>log(NCI)</td>
<td>-0.304$^c$</td>
<td>0.027$^c$</td>
<td>0.068$^c$</td>
<td>-0.414$^c$</td>
</tr>
<tr>
<td></td>
<td>(0.024)</td>
<td>(0.017)</td>
<td>(0.013)</td>
<td>(0.025)</td>
</tr>
<tr>
<td>AEX premium</td>
<td>-0.013</td>
<td></td>
<td>-0.010</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.020)</td>
<td></td>
<td>(0.017)</td>
<td></td>
</tr>
<tr>
<td>Time ($t$)</td>
<td>0.019$^c$</td>
<td></td>
<td>0.022$^c$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.001)</td>
<td></td>
<td>(0.001)</td>
<td></td>
</tr>
<tr>
<td>Retirement dummy</td>
<td>0.069$^c$</td>
<td></td>
<td>0.052$^c$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.018)</td>
<td></td>
<td>(0.015)</td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.290$^c$</td>
<td></td>
<td>0.266$^c$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.004)</td>
<td></td>
<td>(0.003)</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Group 1, $T_i \geq 10$</th>
<th>Group 2, $T_i \geq 10$</th>
<th>Group 3, $T_i \geq 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1, $T_i \geq 9$</td>
<td>0.857</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group 2, $T_i \geq 9$</td>
<td>0.045</td>
<td>0.870</td>
<td>0</td>
</tr>
<tr>
<td>Group 3, $T_i \geq 9$</td>
<td>0.098</td>
<td>0.130</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Ratio</th>
<th>Group 1, $T_i \geq 9$</th>
<th>Group 2, $T_i \geq 9$</th>
<th>Group 3, $T_i \geq 9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1, $T_i \geq 8$</td>
<td>0.674</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Group 2, $T_i \geq 8$</td>
<td>0.109</td>
<td>1.000</td>
<td>0.067</td>
</tr>
<tr>
<td>Group 3, $T_i \geq 8$</td>
<td>0.217</td>
<td>0</td>
<td>0.933</td>
</tr>
</tbody>
</table>

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 $N^{-1}\tilde{\beta}\tilde{\beta}'$ 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. It is possible to extend our method to nonstationary panels with latent group structures. Fourth, it is also possible to allow for structural changes in the model; see, e.g., Okui and Wang (2017). We leave these topics for future research.

REFERENCES


A Some technical lemmas

In this section, we state and prove several technical lemmas that are used in the proofs of the main results in the paper. Section B provides the proofs of the main results in the paper. Sections C and D provide the details of SBSA 2 and some additional simulation results, respectively.

Lemma A.1. Let \( \{w_{it}, t \geq 1\} \) be as defined in Assumption A1(i). Let \( \xi(w_{it}; \varsigma) \) be a \( \mathbb{R}^{d_{\epsilon}} \)-valued function indexed by the parameter \( \varsigma \in \Upsilon \), where \( \Upsilon \) is a convex compact set in \( \mathbb{R}^{d_{\epsilon}} \), \( d_{\epsilon} = O(1) \), \( d_{\epsilon}/T)(\log T)^3 = o(1) \), and \( \mathbb{E}[\xi(w_{it}; \varsigma)] = 0 \) for all \( i = 1, \ldots, N \), \( t = 1, \ldots, T \), and \( \varsigma \in \Upsilon \). Assume that there exists a function \( M(w_{it}) \) such that \( \|\xi(w_{it}; \varsigma) - \xi(w_{it}; \varsigma')\| \leq M(w_{it})\|\varsigma - \varsigma'\| \) for all \( \varsigma, \varsigma' \in \Upsilon \) and \( \sup_{\varsigma} \|\xi(w_{it}; \varsigma)\| \leq M(w_{it}) \).

Assume that \( \mathbb{E}[M(w_{it})]^k < \infty \) for some \( k \geq 6 \) such that \( N = O(T^{k/2-1}) \). Let \( \{\varsigma_i\} \) be a nonstochastic sequence in \( \Upsilon \). Then \( \max_{T-1/2} \sum_{t=1}^{T} \xi(w_{it}; \varsigma_i) = O_P((\ln T)^3) \).

Proof: The proof is essentially identical to that of Lemma S1.2(i) in Su, Shi, and Phillips (2016, SSP hereafter) despite the fact that we now allow \( d_{\epsilon} \) to diverge to infinity.

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

Lemma A.2. Suppose that Assumption A1 holds. Then for any fixed \( \eta > 0 \) and \( v > 0 \), we have

(i) \( P \left( \max_i \sup_{(\gamma_i, \theta)} \left| \Psi_i (\gamma_i, \theta) - \Psi_i (\gamma_i, \theta) \right| \geq \eta \right) = o(N^{-1}) \),

(ii) \( \|\tilde{\gamma}_i (\theta) - \gamma_i (\theta)\| = O_P ((T/p)^{-1/2}) \) for each \( i \),

(iii) \( P \left( \max_i \sup_{\theta} \|\tilde{\gamma}_i (\theta) - \gamma_i (\theta)\| \geq \eta (T/p)^{-1/2} (\ln T)^{3+v} \right) = o(N^{-1}) \),

(iv) \( P \left( \max_i \sup_{\theta} \left| \Psi_i (\tilde{\gamma}_i (\theta), \theta) - \Psi_i (\gamma_i (\theta), \theta) \right| \right) \geq \eta (T/p)^{-1/2} (\ln T)^{3+v} \right) = o(N^{-1}) \),

(v) \( \frac{1}{T} \sum_{t=1}^{T} \|\tilde{\gamma}_i (\theta^0) - \gamma_i (\theta^0)\|^2 = O_P (T^{-1}) \).

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

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

and \( \max_{i,t} \mathbb{E} [M (w_{it})] \leq c_M^{1/k} \) by Assumption A1(iv) and the Jensen inequality, (iv) follows from (iii). We are left to show (v). Recall that \( Z (w_{it}; \gamma_i, \theta) = \partial \varphi (w_{it}; \gamma_i, \theta)/\partial \gamma_i \) and \( Z^{\gamma_i} (w_{it}; \gamma_i, \theta) = \partial Z (w_{it}; \gamma_i, \theta)/\partial \gamma_i \).
Noting that \( \hat{\gamma}_i (\theta) = \arg \min_{\gamma_i} \hat{\Psi}_i (\gamma_i, \theta) \), we have
\[
0 = \frac{1}{T} \sum_{t=1}^{T} Z (w_{it}; \hat{\gamma}_i (\theta), \theta) = \frac{1}{T} \sum_{t=1}^{T} Z (w_{it}; \gamma_i (\theta), \theta) + \frac{1}{T} \sum_{t=1}^{T} Z (w_{it}; \hat{\gamma}_i (\theta), \theta) [\hat{\gamma}_i (\theta) - \gamma_i (\theta)]
\]
where \( \tilde{H}_i (\theta) = \frac{1}{T} \sum_{t=1}^{T} Z (w_{it}; \gamma_i (\theta), \theta) \) and \( \hat{\gamma}_i (\theta) \) lies between \( \hat{\gamma}_i (\theta) \) and \( \gamma_i (\theta) \) elementwise. Then
\[
\hat{\gamma}_i (\theta^0) - \gamma_i (\theta^0) = -\tilde{H}_i (\theta^0) \frac{1}{T} \sum_{t=1}^{T} Z (w_{it}; \gamma_i (\theta^0), \theta^0)
\]
provided that \( \tilde{H}_i (\theta^0) \) is asymptotically nonsingular. Let \( \tilde{H}_i = \hat{H}_i (\theta^0), \tilde{H}_i = \frac{1}{T} \sum_{t=1}^{T} Z (w_{it}; \gamma_i (\theta^0), \theta^0), \) and \( H_i = \mathbb{E}(\tilde{H}_i) \). Under Assumption A1, we can readily show that
\[
\max_{1 \leq i \leq N} \| \tilde{H}_i - H_i \| \leq \max_{1 \leq i \leq N} \| \tilde{H}_i - \hat{H}_i \| + \max_{1 \leq i \leq N} \| \hat{H}_i - H_i \| = o_P (1),
\]
which implies that \( \lambda_{\min}(\tilde{H}_i) = \lambda_{\min}(H_i) + o_P (1) \) uniformly in \( i \). Consequently, we have
\[
\frac{1}{N} \sum_{i=1}^{N} \| \hat{\gamma}_i (\theta^0) - \gamma_i (\theta^0) \|^2 \leq \left[ \min_{1 \leq i \leq N} \lambda_{\min}(\tilde{H}_i) + o_P (1) \right]^{-2} \frac{1}{NT^2} \sum_{i=1}^{N} \| \sum_{t=1}^{T} Z (w_{it}; \gamma_i (\theta^0), \theta^0) \|^2
\]
where we use the fact that \( \mathbb{E}[Z (w_{it}; \gamma_i (\theta^0), \theta^0)] = 0 \) and that \( \mathbb{E}[\| \sum_{t=1}^{T} Z (w_{it}; \gamma_i (\theta^0), \theta^0) \|^2 = O(pT) \) by a simple application of the Davydov inequality under Assumption A1 (e.g., Hall and Heyde (1980, p. 278)).

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

**Proof:** Noting that \( \tilde{\theta} = \arg \min_{\theta} Q_{NT} (\theta) \), we have \( Q_{NT} (\tilde{\theta}) \leq Q_{NT} (\theta^0) \). By Assumption A(iv), there exists a constant \( \epsilon > 0 \) such that \( \inf_{\| \theta - \theta^0 \| > \eta} \frac{1}{N} \sum_{i=1}^{N} \| \tilde{\Psi}_i (\gamma_i (\theta), \theta) - \Psi_i (\gamma_i (\theta^0), \theta^0) \| \geq \epsilon \). Define
\[
A_1 \equiv \left\{ \max_{1 \leq i \leq N} \sup_{\gamma_i (\theta)} | \tilde{\Psi}_i (\gamma_i (\theta), \theta) - \Psi_i (\gamma_i (\theta), \theta) | \leq \frac{1}{6} \right\}, \text{ and}
\]
\[
A_2 \equiv \left\{ \sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{i=1}^{N} \| \tilde{\Psi}_i (\gamma_i (\theta), \theta) - \Psi_i (\gamma_i (\theta), \theta) \| \leq \frac{1}{6} \right\}.
\]
By Lemma A.2(i) and (iii)–(iv) and Assumption A1(ii), \( P(A_1 \cap A_2) \geq 1 - P(A_1^c) - P(A_2^c) = 1 - o(N^{-1}) \).
Then conditional on \( A_1 \cap A_2 \), we have
\[
\inf_{\theta : \| \theta - \theta^0 \| > \eta} \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta) , \theta) \geq \inf_{\theta : \| \theta - \theta^0 \| > \eta} \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta) , \theta) - \frac{1}{6} \\
\geq \inf_{\theta : \| \theta - \theta^0 \| > \eta} \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta^0) , \theta^0) + \epsilon - \frac{1}{6} \epsilon - \frac{1}{6} \epsilon \\
\geq \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta^0) , \theta^0) - \frac{1}{6} \epsilon + \epsilon - \frac{1}{6} \epsilon - \frac{1}{6} \epsilon \\
\geq \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta^0) , \theta^0) - \frac{1}{6} \epsilon - \frac{1}{6} \epsilon + \epsilon - \frac{1}{6} \epsilon - \frac{1}{6} \epsilon \\
= \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta^0) , \theta^0) + \frac{1}{3} \epsilon.
\]

On the other hand, \( \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\bar{\theta}) , \bar{\theta}) \leq \frac{1}{N} \sum_{i=1}^{N} \Psi_i (\gamma_i (\theta^0) , \theta^0) \). It follows that \( P(|| \bar{\theta} - \theta^0 || > \eta) = o(N^{-1}) \).

To state and prove the next lemma, we follow Hahn and Newey (2004) and SSP and introduce some notation. Let \( F_i \) and \( \hat{F}_i \) denote the cumulative and empirical distribution functions of \( w_{it} \), respectively. Let \( F_i (\epsilon) = F_i + \epsilon \sqrt{T} (\hat{F}_i - F_i) \) for \( \epsilon \in [0,T^{-1/2}] \). For fixed \( \theta \) and \( \epsilon \), let \( \gamma_i (\theta, F_i (\epsilon)) = \arg \min_{\gamma_i} \int \psi (\cdot ; \gamma_i, \theta) dF_i (\epsilon) \), which is the solution to the estimating equation
\[
0 = \frac{1}{N} \sum_{i=1}^{N} \int Z_i (\cdot ; \gamma_i (\theta, F_i (\epsilon)) , \theta) dF_i (\epsilon).
\]

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

**Lemma A.4.** Suppose that Assumption A1 holds. Then
\( (i) \) \( P (\max_{1 \leq i \leq N \max_{0 \leq \epsilon \leq T^{-1/2}}} \| \gamma_i (\theta, F_i (\epsilon)) - \gamma_i (\theta) \| \geq \eta) = o(N^{-1}) \) for any \( \eta > 0 \),
\( (ii) \) \( \max_{1 \leq i \leq N \max_{0 \leq \epsilon \leq T^{-1/2}}} \| \gamma_i (\theta) - \gamma_i (\theta^0) \| = o(1) \),
\( (iii) \) \( P (\max_{1 \leq i \leq N \max_{0 \leq \epsilon \leq T^{-1/2}}} \| \gamma_i (\theta) - \bar{\gamma}_i (\theta^0) \| \geq \eta) = o(N^{-1}) \) for any \( \eta > 0 \),
\( (iv) \) \( P (\max_{1 \leq i \leq N \max_{0 \leq \epsilon \leq T^{-1/2}}} \| \frac{\partial \gamma_i (\theta, F_i (\epsilon))}{\partial \theta^T} - \frac{\partial \gamma_i (\theta^0)}{\partial \theta^T} \| \geq \eta) = o(N^{-1}) \) for any \( \eta > 0 \),
\( (v) \) \( \max_{1 \leq i \leq N \max_{0 \leq \epsilon \leq T^{-1/2}}} \| \frac{\partial \gamma_i (\theta)}{\partial \theta^T} - \frac{\partial \gamma_i (\theta^0)}{\partial \theta^T} \| = o(1) \),
\( (vi) \) \( P (\max_{1 \leq i \leq N \max_{0 \leq \epsilon \leq T^{-1/2}}} \| \frac{\partial \gamma_i (\theta)}{\partial \theta^T} - \frac{\partial \gamma_i (\theta^0)}{\partial \theta^T} \| \geq \eta) = o(N^{-1}) \) for any \( \eta > 0 \).

**Proof:** The proofs of (i)–(iii) parallel those of Lemma S1.8(i)–(iii) in SSP and thus are omitted. Similarly, the proofs of (iv)–(vi) parallel those of Lemma S1.9(i)–(iii) in SSP. ■
Proof of Theorem 3.1: (i) Noting that \( \hat{\theta} = \arg \min_\theta Q_{NT}(\theta) \), by the second order Taylor expansion and the envelope theorem, we have

\[
0 \geq Q_{NT}(\hat{\theta}) - Q_{NT}(\theta^0) = \frac{1}{N} \sum_{i=1}^{N} \left[ \hat{\Psi}_i(\tilde{\gamma}_i(\hat{\theta}), \hat{\theta}) - \hat{\Psi}_i(\tilde{\gamma}_i(\theta^0), \theta^0) \right]
\]

\[
= \delta^T \hat{S} + \frac{1}{2} \delta^T \hat{H}(\hat{\theta}) \delta \geq \frac{1}{2} \lambda_{\min}(\hat{H}(\hat{\theta})) \|\delta\|^2 - \|\hat{S}\| \cdot \|\delta\|,
\]

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

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

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

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

Noting that \( \mathbb{E}(S_1) = 0 \) and \( \text{tr}[\text{Var}(S_1)] = O(q(NT)^{-1}) \), we have \( \|S_1\| = O_p((NT/q)^{-1/2}) \). For \( S_2 \), we have by the Cauchy-Schwarz and Markov inequalities, Assumption A1(iv), and Lemma A.2(v)

\[
\|S_2\| \leq \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \|W^{\gamma_i}(w_{it}; \gamma_i(\theta^0), \theta) \cdot \|\tilde{\gamma}_i(\theta) - \gamma_i(\theta^0)\|\]

\[
\leq \left[ \frac{1}{N} \sum_{i=1}^{N} \left( \frac{1}{T} \sum_{t=1}^{T} M(w_{it})^2 \right)^{1/2} \right]^{1/2} \left[ \frac{1}{N} \sum_{i=1}^{N} \|\tilde{\gamma}_i(\theta^0) - \gamma_i(\theta^0)\|^2 \right]^{1/2}
\]

\[
= O_p(1) O_p((q/T)^{1/2}) = O_p((q/T)^{1/2}).
\]

Then \( \|\hat{S}\| = O_p((q/T)^{1/2}) \).

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

\[
\|\hat{H}(\hat{\theta}) - H(\theta^0)\| \leq \|\hat{H}(\hat{\theta}) - \hat{H}(\theta^0)\| + \|\hat{H}(\theta^0) - \hat{H}(\theta^0)\| + \|\hat{H}(\theta^0) - H(\theta^0)\|.
\]

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

(ii) Let \( a_T = (p/T)^{1/2} \) and \( v \in \mathbb{R}^{p+r} \). Following Lam and Fan (2008) and Qian and Su (2016), we want to show that for any \( \epsilon > 0 \), there is a constant \( C \) such that

\[
P \left( \inf_{\|v\|=C} \hat{\Psi}_i(\gamma_i^0 + a_T v, \hat{\theta}) - \hat{\Psi}_i(\gamma_i^0, \hat{\theta}) \geq 1 - \epsilon \right) \geq 1 - \epsilon.
\]

This implies that w.p.a.1, \( \hat{\Psi}_i(\cdot, \hat{\theta}) \) has a local minimum at \( \tilde{\gamma}_i \), where \( \tilde{\gamma}_i \in \{ \gamma_i^0 + a_T v : \|v\| \leq C \} \). So \( \|\tilde{\gamma}_i - \gamma_i^0\| = O_P(a_T) \). Let

\[
\Xi_i(v) = \hat{\Psi}_i(\gamma_i^0 + a_T v, \hat{\theta}) - \hat{\Psi}_i(\gamma_i^0, \hat{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \left[ \varphi(w_{it}; \gamma_i^0 + a_T v, \hat{\theta}) - \varphi(w_{it}; \gamma_i^0, \hat{\theta}) \right]
\]

\[
= a_T v^T \hat{S}_{zi} + \frac{1}{2} a_T v^T \hat{H}_{i,\theta\theta}(\gamma_i^0, \hat{\theta}) v \equiv \Xi_{i,1}(v) + \Xi_{i,2}(v),
\]
where \( \tilde{\gamma}_i \) lies between \( \gamma_i^0 + aTv \) and \( \gamma_i^0 + \tilde{S}_Z \), \( \tilde{S}_Z = \frac{1}{T} \sum_{t=1}^T Z(w_{it}; \gamma_i^0, \tilde{\theta}) \), and \( \tilde{H}_{i,\gamma}(\gamma_i, \theta) = \frac{1}{T} \sum_{t=1}^T Z(\gamma_i; \gamma_i, \theta) \). As in the proof of Lemma A.2(v) and using the result in part (i), we can readily show that under Assumption A1, \( \|\tilde{S}_Z\| = \|\frac{1}{T} \sum_{t=1}^T Z(w_{it}; \gamma_i^0, \tilde{\theta}) + \mu \| = O_p((p/T)^{1/2}) \) and \( \lambda_{\min}(\tilde{H}_{i,\gamma}(\gamma_i, \tilde{\theta})) = \lambda_{\min}(H_{i,\gamma}(\theta^0)) + O_p(1) \) uniformly in \( i \). Then by Chebyshev inequality and Assumption A1(v), we have that

\[
\|\tilde{\gamma}_i - \gamma_i^0\| = O_p(aTv).
\]

(iii) By Lemma A.1 and the result in (i), we can show that \( \gamma_i \) is a permutation of \( \{1, 2, \ldots, N\} \) and \( \gamma_i \) is satisﬁed for \( i = 1, \ldots, N \) that is implicitly determined by the ranking relation in (B.1).

**Proof of Theorem 3.2:** Let \( u_i = \tilde{\beta}_i - \beta_i^0 \). By Theorem 3.1 and Assumption A1(vi), \( \|u_i\| = O_p((p/T)^{1/2}) \) and \( \max_{1 \leq i \leq N} \|u_i\| = O_p(p/T)^{1/2}((\ln T)^3) = o_P(1) \). Without loss of generality (W.l.o.g.), we focus on the proof of the theorem when \( K = 0 \) and then remark on the other cases. By ranking the preliminary estimates \( \{\tilde{\beta}_i\} \) according to their \( j \)-th elements, we have

\[
\tilde{\beta}_{\pi_j(1),j} \leq \tilde{\beta}_{\pi_j(2),j} \leq \cdots \leq \tilde{\beta}_{\pi_j(N),j} \quad \text{for} \quad j = 1, \ldots, p,
\]

(B.1)

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

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

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

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

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

Let \( S_l \) denote the collection of the regressor indices such that the conditions in Case \( l \) are satisﬁed for \( l = 1, 2, 3 \). Apparently, \( S_1 \cup S_2 \cup S_3 = \{1, 2, \ldots, p\} \) and \( S_l \cap S_{l'} = \emptyset \) for \( l \neq l' \). Assumption A2(i) ensures that \( S_1 \cup S_2 \) must be nonempty.

Let \( j_1 = \arg \max_{1 \leq j \leq p} \hat{V}_{1,N}(j) \) where \( \hat{V}_{1,N}(j) = \hat{V}_{1,N}^0(j)/\sigma_{1,N}^2(j) \). Apparently, \( \sigma_{1,N}^2(j) \) is bounded away from zero in probability for each \( j \). By Theorem 3.1, the sample variance of \( \{\tilde{\beta}_{ij}, i = 1, \ldots, N\} \) converges to a positive constant \( c_j \), say, for any \( j \in S_1 \cup S_2 \), whereas that of \( \{\tilde{\beta}_{ij}, i = 1, \ldots, N\} \) is \( O(p/T) \) for any \( j' \in S_3 \). As a result, \( P(j_1 \in S_1 \cup S_2) \rightarrow 1 \) and index \( j_1 \) is chosen to estimate the break points in the whole sample \( S_{1,N}(j) = \{\tilde{\beta}_{\pi_{j_1}(1),j}, \tilde{\beta}_{\pi_{j_2}(2),j}, \ldots, \tilde{\beta}_{\pi_{j_N}(N),j}\} \) in the first step of the SBSA for some \( j \in S_1 \cup S_2 \) such
that \( j_1 = j \). We will show no matter whether \( j \) is in \( S_1 \) or \( S_2 \), we can always identify one break point in \( S_{1,N}(j) \) w.p.a.1.

The second break point can be identified by choosing \( j \in S_1 \) or \( j \in S_1 \cup S_2 \) depending on whether the break point \( \{N_1 + N_2\} \) or \( \{N_1\} \) is identified in the first step. For example, if in the first step we rely on some \( j \in S_1 \) to identify the break point \( \{N_1\} \) that distinguishes the first group from the rest two groups, then in the second step, we may rely on an element \( j \) from either \( S_1 \) or \( S_2 \) to identify the second break point \( \{N_1 + N_2\} \) that separates the second group from the third group. On the other hand, if the break point \( \{N_1 + N_2\} \) is identified in the first step to separate the third group from the rest two groups, then in the second step, we can only rely on some \( j \in S_1 \) to identify the break point \( \{N_1\} \) to separate the first and second groups. In this second case, we will show that \( P(\hat{m}_1 = N_1 + N_2) \to 1 \) as \((N, T) \to \infty\), which implies that w.p.a.1 \( j_2 \equiv \arg \max_{1 \leq j \leq p}[\hat{V}_{1,N_1} + \hat{V}_{N_1 + N_2}(j)] = \arg \max_{1 \leq j \leq p}[\hat{V}_{1,N_1 + N_2}(j) + \hat{V}_{N_1 + N_2 + 1, N}(j)] \). Since the segment \( S_{N_1 + N_2 + 1, N}(j) \) does not contain any break point, \( \hat{V}_{N_1 + N_2 + 1, N}(j) = O_P(p/T) \) for any \( j \in \{1, 2, \ldots, p\} \) by Theorem 3.1. But \( \hat{V}_{1,N_1 + N_2}(j) \) is bounded away from zero in probability for any \( j \in S_1 \) and \( O_P(p/T) \) for any \( j \in S_2 \cup S_3 \). As a result, \( P(j_2 \in S_1) \to 1 \) as \((N, T) \to \infty\). Our choice of selecting \( j_{k-1} \) in the SBSA ensures that such an argument continues to hold when we have \( K^0 > 3 \) groups as long as the \( K^0 \) groups are separable from each other as required explicitly in Assumption A2(i).

Below, we prove that when either Case 1 or Case 2 applies (i.e., \( j_1 \in S_1 \) or \( j_1 \in S_2 \), we can successfully identify one break point in the first step of the SBSA. After one break point is identified in the first step, we can also identify the second break point in the second step no matter whether \( j_2 \in S_1 \) or \( j_2 \in S_2 \).

**Case 1**: \( j_1 \in S_1 \). Based on the ranking relation in (B.1) and the fact that \( \max_{1 \leq i \leq N} \|u_i\| = o_P(1) \), we have the following homogeneity property

\[
\beta^0_{\tau_i(j)} = \begin{cases} 
\alpha_1^0 & \text{if } 1 \leq i \leq N_1, \\
\alpha_2^0 & \text{if } N_1 + 1 \leq i \leq N_1 + N_2, \text{ for any } j \in S_1, \\
\alpha_3^0 & \text{if } N_1 + N_2 + 1 \leq i \leq N,
\end{cases}
\]

Fix \( j \in S_1 \) and \( \hat{m}_1(j) = \arg \min_{1 \leq m < N} \hat{S}_{1,N}(j, m) \). We consider three subcases:

**Case 1a**: \( \frac{\tau_1}{\tau_1 + \tau_2}(\alpha_1^0 - \alpha_2^0)^2 > \frac{\tau_3}{\tau_2 + \tau_3}(\alpha_2^0 - \alpha_3^0)^2 \), ensuring \( P\left(\hat{S}_{1,N}(j, N_1 + N_2) - \hat{S}_{1,N}(j, N_1) > 0\right) \to 1 \).

**Case 1b**: \( \frac{\tau_1}{\tau_1 + \tau_2}(\alpha_1^0 - \alpha_2^0)^2 > \frac{\tau_3}{\tau_2 + \tau_3}(\alpha_2^0 - \alpha_3^0)^2 \), ensuring \( P\left(\hat{S}_{1,N}(j, N_1 + N_2) - \hat{S}_{1,N}(j, N_1) < 0\right) \to 1 \).

**Case 1c**: \( \frac{\tau_1}{\tau_1 + \tau_2}(\alpha_1^0 - \alpha_2^0)^2 = \frac{\tau_3}{\tau_2 + \tau_3}(\alpha_2^0 - \alpha_3^0)^2 \).

W.l.o.g., we focus on Case 1a and will show that \( P(\hat{m}_1(j) = N_1) \to 1 \) as \((N, T) \to \infty\) by proving that (i) \( P(\hat{m}_1(j) < N_1) \to 0 \); (ii) \( P(N_1 < \hat{m}_1(j) \leq N_1 + N_2) \to 0 \); (iii) \( P(N_1 + N_2 < \hat{m}_1(j) \leq N) \to 0 \) as \((N, T) \to \infty\). Then by mere symmetry, we can show that \( P(\hat{m}_1(j) = N_1 + N_2) \to 1 \) as \((N, T) \to \infty\) in Case 1b, and by arguments similar to those used in the proof of Lemma 7 in Bai (1997), we can show that \( P(\hat{m}_1(j) = N_1) = P(\hat{m}_1(j) = N_1 + N_2) = \frac{1}{2} \). In each subcase, we can identify one break point \( \{N_1\} \) or \( \{N_1 + N_2\} \) in the first step of the SBSA.

We first show (i). When \( m < N_1 \), the average of \( \tilde{\beta}_{1,j} \) over the two binary segments are

\[
\tilde{\beta}_{1,m}(j) = \alpha_{1,j}^0 + \bar{u}_{1,m}(j) \quad \text{and} \quad \tilde{\beta}_{m+1,N}(j) = \frac{N_1 - m}{N - m}\alpha_{1,j}^0 + \frac{N_2}{N - m}\alpha_{2,j}^0 + \frac{N_3}{N - m}\alpha_{3,j}^0 + \bar{u}_{m+1,N}(j),
\]

\[\text{The condition can also be written as } \frac{\tau_1}{\tau_1 + \tau_2}(\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 \geq \frac{1 - \tau_1}{1 - \tau_1}(\alpha_{2,j}^0 - \alpha_{3,j}^0)^2, \text{ similar to equation (6) in Bai (1997).}\]
where $\bar{u}_{1,m}(j) = m^{-1} \sum_{i=1}^{m} u_{\pi_j(i),j}$ and $\bar{u}_{m+1,N}(j) = (N - m)^{-1} \sum_{i=m+1}^{N} u_{\pi_j(i),j}$. Noting that $\hat{\beta}_{\pi_j(i),j} - \hat{\beta}_{1,m}(j) = u_{\pi_j(i),j} - \bar{u}_{1,m}(j)$ when $m < N_1$, we have

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

Similarly, noting that

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

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

$$\sum_{i=m+1}^{N} |\hat{\beta}_{\pi_j(i),j} - \hat{\beta}_{m+1,N}(j)|^2 = (N_1 - m)|a_{1m}|^2 + N_2|a_{2m}|^2 + N_3|a_{3m}|^2$$

$$+ 2a_{1m} \sum_{i=m+1}^{N_1} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)| + 2a_{2m} \sum_{i=N_1+1}^{N_1+N_2} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|$$

$$+ 2a_{3m} \sum_{i=N_1+N_2+1}^{N} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2.$$

It follows that

$$\hat{S}_{1,j}(m) = \frac{1}{N} \left( \sum_{i=1}^{m} |\hat{\beta}_{\pi_j(i),j} - \hat{\beta}_{1,m}(j)|^2 + \sum_{i=m+1}^{N} |\hat{\beta}_{\pi_j(i),j} - \hat{\beta}_{m+1,N}(j)|^2 \right) = M_{1j}(m) + r_{1j}(m),$$

where $M_{1j}(m) = \frac{N_1-m}{N}|a_{1m}|^2 + \frac{N_2}{N}|a_{2m}|^2 + \frac{N_3}{N}|a_{3m}|^2$, and

$$r_{1j}(m) = \frac{1}{N} \left[ \sum_{i=1}^{m} |u_{\pi_j(i),j} - \bar{u}_{1,m}(j)|^2 + \sum_{i=m+1}^{N} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2 \right]$$

$$+ 2a_{1m} \sum_{i=m+1}^{N_1} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)| + 2a_{2m} \sum_{i=N_1+1}^{N_1+N_2} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|$$

$$+ 2a_{3m} \sum_{i=N_1+N_2+1}^{N} |u_{\pi_j(i),j} - \bar{u}_{m+1,N}(j)|^2.$$

Noting that $a_{1m}$, $a_{2m}$, and $a_{3m}$ are $O(1)$ uniformly in $m \leq N_1$, we can readily apply Theorem 3.1(iii) and show that $r_{1j}(m) = O_p((p/T)^{1/2}(\ln T)^3)$ uniformly in $m < N_1$. Now, observe that $\hat{S}_{1,j}(j,N_1) - \hat{S}_{1,j}(j,N_1) = [M_{1j}(m) - M_{1j}(N_1)] + [r_{1j}(m) - r_{1j}(N_1)]$. By straightforward but tedious calculations, we can show that

$$\Delta M_{1j}(m) = M_{1j}(m) - M_{1j}(N_1)$$

$$= \frac{N_1 - m}{N(1 - m/N)(1 - N_1/N)} \left( \left(1 - \frac{N_1}{N}\right)(\alpha^{0}_{1,j} - \alpha^{0}_{2,j}) + \left(1 - \frac{N_1 + N_2}{N}\right)(\alpha^{0}_{2,j} - \alpha^{0}_{3,j}) \right)^2$$

$$\approx \mu_{1mN}$$

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

\[ r_{1j}(m) - r_{1j}(N_1) = \frac{1}{N} \left[ \sum_{i=1}^{m} |u_{\tau_j(i),j} - \bar{u}_{1,m}(j)|^2 - \sum_{i=1}^{N_1} |u_{\tau_j(i),j} - \bar{u}_{1,N_1}(j)|^2 \right] \]

\[ + \frac{1}{N} \left[ \sum_{i=m+1}^{N} |u_{\tau_j(i),j} - \bar{u}_{m+1,N}(j)|^2 - \sum_{i=N_1+1}^{N} |u_{\tau_j(i),j} - \bar{u}_{N_1+1,N}(j)|^2 \right] \]

\[ + \frac{2a_1 m}{N} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j} - \bar{u}_{m+1,N}(j)| \]

\[ + \frac{2}{N} \left\{ a_2 m \sum_{i=N_1+1}^{N_1+N_2} |u_{\tau_j(i),j} - \bar{u}_{m+1,N}(j)| - a_2 N_1 \sum_{i=N_1+1}^{N_1+N_2} |u_{\tau_j(i),j} - \bar{u}_{N_1+1,N}(j)| \right\} \]

\[ + \frac{2}{N} \left\{ a_3 m \sum_{i=N_1+N_2+1}^{N} |u_{\tau_j(i),j} - \bar{u}_{m+1,N}(j)| - a_3 N_1 \sum_{i=N_1+N_2+1}^{N} |u_{\tau_j(i),j} - \bar{u}_{N_1+1,N}(j)| \right\} \]

\[ = I_{1,m} + I_{2,m} + 2I_{3,m} + 2I_{4,m} + 2I_{5,m}, \text{ say.} \]

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

\[ I_{1,m} = -\frac{1}{N} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j}|^2 + \frac{1}{N} \left[ N_1 |\bar{u}_{1,N_1}(j)|^2 - m|\bar{u}_{1,m}(j)|^2 \right] \]

\[ = -\frac{1}{N} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j}|^2 + \frac{N_1-m}{N} |\bar{u}_{1,N_1}(j)|^2 + \frac{m}{N} [\bar{u}_{1,N_1}(j) - \bar{u}_{1,m}(j)] \cdot [\bar{u}_{1,N_1}(j) + \bar{u}_{1,m}(j)] \]

\[ = \frac{N_1-m}{N} \left\{ -\frac{1}{N_1-m} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j}|^2 + |\bar{u}_{1,N_1}(j)|^2 \right\} \]

\[ - \left\{ \frac{1}{N_1-m} \sum_{i=1}^{N_1} u_{\tau_j(i),j} - \frac{1}{N_1-m} \sum_{i=m+1}^{N_1} u_{\tau_j(i),j} \right\} \cdot [\bar{u}_{1,N_1}(j) + \bar{u}_{1,m}(j)] \]

\[ = \mu_{1mN} \cdot O_P(pT^{-1}(\ln T)^6) = o_P(\mu_{1mN}). \]

Similarly, noting that \( \bar{u}_{N_1+1,N}(j) - \bar{u}_{m+1,N}(j) = \frac{1}{N-m} \sum_{i=N_1+1}^{N} u_{\tau_j(i),j} - \frac{1}{N-m} \sum_{i=m+1}^{N} u_{\tau_j(i),j} = \frac{N_1-m}{N-N_1} \times \left[ \frac{1}{N-m} \sum_{i=m+1}^{N} u_{\tau_j(i),j} - \frac{1}{N-m} \sum_{i=m+1}^{N} u_{\tau_j(i),j} \right] \), we have

\[ I_{2,m} = \frac{1}{N} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j}|^2 + \frac{1}{N} \left[ (N-N_1) |\bar{u}_{N_1+1,N}(j)|^2 - (N-m)|\bar{u}_{m+1,N}(j)|^2 \right] \]

\[ = \frac{1}{N} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j}|^2 - \frac{N_1-m}{N} |\bar{u}_{m+1,N}(j)|^2 + \frac{N_1-N_1}{N} [\bar{u}_{N_1+1,N}(j) - \bar{u}_{m+1,N}(j)] \cdot [\bar{u}_{N_1+1,N}(j) + \bar{u}_{m+1,N}(j)] \]

\[ = \frac{N_1-m}{N} \left\{ \frac{1}{N_1-m} \sum_{i=m+1}^{N_1} |u_{\tau_j(i),j}|^2 - |\bar{u}_{m+1,N}(j)|^2 \right\} \]

\[ + \frac{N_1-m}{N} \left\{ \frac{1}{N-m} \sum_{i=m+1}^{N} u_{\tau_j(i),j} - \frac{1}{N_1-m} \sum_{i=m+1}^{N_1} u_{\tau_j(i),j} \right\} \cdot [\bar{u}_{N_1+1,N}(j) + \bar{u}_{m+1,N}(j)] \]

\[ = \mu_{1mN} \cdot O_P(pT^{-1}(\ln T)^6) = o_P(\mu_{1mN}). \]
Thus, noting that

\[ a_3 = 0 \]

we have

\[ I_3 = \frac{N_1 - m}{N} \sum_{i=m+1}^{N_1} u_{\pi_j(i),j} \]

\[ = \mu_{1mN} \cdot O_P((p/T)^{1/2}(ln T)^3) = o_p(\mu_{1mN}). \]

For \( I_{4,m} \), noting that

\[ |a_2 - a_{2N}| = \left| \frac{N_1 - m}{N - m} (\alpha_{2,j}^0 - \alpha_{1,j}^0) + \left( \frac{N_3}{N - m} - \frac{N_3}{N - N_1} \right) (\alpha_{3,j}^0 - \alpha_{2,j}^0) \right| \]

we have

\[ I_{4,m} = \frac{a_{2m}}{N} \sum_{i=N_1+1}^{N_1+N_2} u_{\pi_j(i),j} - \frac{a_{2N}}{N} \sum_{i=N_1+1}^{N_1+N_2} u_{\pi_j(i),j} \]

\[ = (a_3 - a_{2N}) \frac{N_2 a_{2N}}{N} \sum_{i=N_1+1}^{N_1+N_2} u_{\pi_j(i),j} \]

\[ = (a_3 - a_{2N}) \frac{N_2 a_{2N}}{N} \sum_{i=N_1+1}^{N_1+N_2} u_{\pi_j(i),j} \]

\[ = \mu_{1mN} \cdot O_P((p/T)^{1/2}(ln T)^3) + \mu_{1mN} \cdot O_P((p/T)^{1/2}(ln T)^3) = o_p(\mu_{1mN}). \]

Similarly, noting that

\[ |a_3 - a_{3N}| = \left| \frac{N_1 - m}{N - m} (\alpha_{3,j}^0 - \alpha_{1,j}^0) + \left( \frac{N_2}{N - m} - \frac{N_2}{N - N_1} \right) (\alpha_{3,j}^0 - \alpha_{2,j}^0) \right| \]

we can show that

\[ I_{5,m} = \frac{a_{3m}}{N} \sum_{i=N_1+N_2+1}^{N} u_{\pi_j(i),j} - \frac{a_{3N}}{N} \sum_{i=N_1+N_2+1}^{N} u_{\pi_j(i),j} \]

\[ = (a_3 - a_{3N}) \frac{N_3 a_{3N}}{N} \sum_{i=N_1+N_2+1}^{N} u_{\pi_j(i),j} \]

\[ = \mu_{1mN} \cdot O_P((p/T)^{1/2}(ln T)^3) + \mu_{1mN} \cdot O_P((p/T)^{1/2}(ln T)^3) = o_p(\mu_{1mN}). \]

Thus \( r_1(m) - r_1(N_1) = o_p(\mu_{1mN}) \) uniformly in \( m < N_1 \) and

\[ P(\tilde{m}_1(j) < N_1) \leq P(\exists m < N_1, \tilde{S}_{1,N}(j,m) - \tilde{S}_{1,N}(j,N_1) < 0) \]

\[ = P(\exists m < N_1, \Delta M_1(m) + [r_1(m) - r_1(N_1)] < 0) \rightarrow 0 \]
as \((N, T) \to \infty\). Then (i) follows.

We now study case (ii). Noting that when \(N_1 < m \leq N_1 + N_2\), \(\tilde{\beta}_{1,m}(j) = \frac{N_1}{m} \alpha_{0,j}^0 + \frac{m - N_1}{m} \alpha_{2,j}^0 + \tilde{u}_{1,m}(j)\), and \(\tilde{\beta}_{m+1,N}(j) = \frac{N_1 + N_2 - m}{N - m} \alpha_{0,j}^0 + \frac{m}{N - m} \alpha_{2,j}^0 + \tilde{u}_{m+1,N}(j)\). It follows that for \(i = 1, \ldots, m\),

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

where \(b_{1m} = \frac{m - N_2}{m} (\alpha_{0,j}^0 - \alpha_{2,j}^0), b_{2m} = \frac{N_2}{m} (\alpha_{2,j}^0 - \alpha_{1,j}^0), \) and we suppress their dependence on \(j\) for notational simplicity. So for the left segment, we have

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

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

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

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

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

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

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

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

$$\Delta M_{2j}(m) = M_{2j}(m) - M_{2j}(N_1) = \frac{N_1}{N} |b_{1m}|^2 + \frac{m-N_1}{N} |b_2m|^2 + \frac{N_1+N_2-m}{N} |b_3m|^2 + \frac{N_2}{N} |b_4m|^2 - \frac{N_2}{N} |b_{3N_1}|^2 - \frac{N_2}{N} |b_{4N_1}|^2$$

$$= \frac{m-N_1}{N} \left[ \frac{N_1}{m} (\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{N_2^2}{(N-N_1)(N-m)} (\alpha_{2,j}^0 - \alpha_{3,j}^0)^2 \right]$$

$$\geq \frac{m-N_1}{N} \frac{N_1+N_2}{m} \left[ \frac{N_1}{N_1+N_2} (\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{N_3}{N-N_1} (\alpha_{2,j}^0 - \alpha_{3,j}^0)^2 \right]$$

$$= \mu_{2mN} \frac{N_1+N_2}{m} \left[ \frac{\tau_1}{\tau_1+\tau_2} (\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{\tau_3}{\tau_2+\tau_3} (\alpha_{2,j}^0 - \alpha_{3,j}^0)^2 \right] \cdot [1 + o(1)]$$

$$\approx \mu_{2mN} \text{ uniformly in } m \in \{N_1 + 1, \ldots, N_1 + N_2\},$$

where the inequality follows from the fact that $\frac{N_1}{N-m} \leq 1 \leq \frac{N_1 + N_2}{m}$ for all $m \in \{N_1 + 1, \ldots, N_1 + N_2\}$, $\mu_{2mN} = \frac{N_2 - N_1}{N-m}$, and the last line follows from the fact that $\frac{\tau_1}{\tau_1+\tau_2} (\alpha_{1,j}^0 - \alpha_{2,j}^0)^2 - \frac{\tau_3}{\tau_2+\tau_3} (\alpha_{2,j}^0 - \alpha_{3,j}^0)^2 > 0$ in Case 1a. Following the analysis of $r_{1j}(m) - r_{1j}(N_1)$, we can show that $r_{2j}(m) - r_{2j}(N_1) = o_P(\mu_{2mN})$ uniformly in $m \in \{N_1 + 1, \ldots, N_1 + N_2\}$. It follows that as $(N, T) \to \infty$, for any $j \in S_2$ we must have

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

Analogously, we can show (iii). It follows that $P(\hat{m}_1(j) = N_1) \to 1$ as $(N, T) \to \infty$ in Case 1a. In other words, by using the ranking relation (B.1) based on regressor $j_1 = j \in S_1$, we could find the right break point w.p.a.1. in the first round of the SBSA. For the ease of presentation, we continue to use $j_1$ to represent $j_1$. Given the first identified break point being $\{N_1\}$ in Case 1a, we have

$$\hat{S}_{1,N}(j, N_1) = \frac{1}{N} \sum_{1 \leq i \leq N_1} \left| \hat{\beta}_{j_1(i), j} - \hat{\beta}_{1,N_1}(j) \right|^2 + \sum_{N_1+1 \leq i \leq N} \left| \hat{\beta}_{j_1(i), j} - \hat{\beta}_{N_1+1,N}(j) \right|^2$$

$$= \frac{1}{N} \left[ N_2 \left( \alpha_{2,j}^0 - \frac{N_2}{N_2 + N_3} \alpha_{2,j}^0 - \frac{N_3}{N_2 + N_3} \alpha_{3,j}^0 \right)^2 + N_3 \left( \alpha_{3,j}^0 - \frac{N_2}{N_2 + N_3} \alpha_{2,j}^0 - \frac{N_3}{N_2 + N_3} \alpha_{3,j}^0 \right)^2 \right]$$

$$+ o_P(1)$$

$$= \frac{1}{N} \frac{N_2 N_3}{N_2 + N_3} \left( \alpha_{2,j}^0 - \alpha_{3,j}^0 \right)^2 + o_P(1) \frac{\tau_2 \tau_3 c_{23,j}^0}{\tau_2 + \tau_3} \equiv \Delta_{1,j},$$

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

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

In the second step, we repeat the iterative algorithm on $\hat{G}_1(2)$ and $\hat{G}_2(2)$. When $j_2 \in S_1$, and $\{N_1\}$ is identified in the first step w.p.a.1. $j_2$ can belong to either $S_1$ or $S_2$. We show that no matter which value $j_2$ takes in $S_1 \cup S_2$, we can identify the second break point $\{N_1 + N_2\}$ w.p.a.1. For the binary segments over $\hat{G}_2(2)$, following similar arguments leading to $P(\hat{m}_2(j) = N_1) \to 1$ as $(N, T) \to \infty$, we can show that $P(\hat{m}_2(j) = N_2) \to 1$ for any $j \in S_1 \cup S_2$ (and thus also for $j_2$) where $\hat{m}_2(j) = \arg \min_{1 \leq m < N_2 + N_3} \hat{S}_{1,N_1+1,N}(j, m)$. Then $\hat{G}_2(2)$ is divided into two sub-segments $\hat{G}_{21}(2)$ and $\hat{G}_{22}(2)$ such that

$$P \left( \hat{G}_{21}(2) = \{\pi_j(N_1 + 1), \ldots, \pi_j(N_1 + N_2)\}, \hat{G}_{22}(2) = \{\pi_j(N_1 + N_2 + 1), \ldots, \pi_j(N)\} \right) \to 1.$$
Furthermore, we can show that by using arguments used in the proof of Theorem 3.1 w.p.a.1. When

This gives 3 groups with

Then based on our SBSA, matter what value

relation in (B.1). Now based on (B.1) and the fact that

w.p.a.1. When

Recall that

Case 2: \(j_1 \in S_2\). W.l.o.g., we consider \(\alpha_{0,j}^0 = \alpha_{2,j}^0 < \alpha_{3,j}^0\) for \(j_1 = j \in S_2\). In this case, it is impossible to distinguish elements from Group 1 from those from Group 2 according to the regressor-\(j\)-based ranking relation in (B.1). Now based on (B.1) and the fact that \(\max_{1 \leq i \leq N} \|u_i\| = O_P((p/T)^{1/2}(\ln T)^3)\), we have the following homogeneity property

\[
\beta_{\pi_j(i),j}^0 = \begin{cases} 
\alpha_{1,j}^0 = \alpha_{2,j}^0 & 1 \leq i \leq N_1 + N_2, \\
\alpha_{3,j}^0 & N_1 + N_2 + 1 \leq i \leq N.
\end{cases}
\]

Recall that \(\hat{m}_1 = \arg \min_{1 \leq m \leq N} \hat{S}_{1,j}(N,j,m)\) where we suppress the dependence of \(\hat{m}_1\) on \(j\). As in Case 1, we want to show as \((N,T) \to \infty\), \(P(\hat{m}_1 = N_1 + N_2) \to 1\) for any \(j \in S_2\) by showing that (i1) \(P(\hat{m}_1 < N_1 + N_2) \to 0\) and (i2) \(P(\hat{m}_1 > N_1 + N_2) \to 0\).

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

\[
\sum_{i=1}^{m} |\tilde{\beta}_{\pi_j(i),j} - \tilde{\beta}_{1,m}(j)|^2 = \sum_{i=1}^{m} |u_{\pi_j(i),j} - \tilde{u}_{1,m}(j)|^2.
\]
Similarly, for \( i = m + 1, \ldots, N \), we have

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

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

\[
\sum_{i=m+1}^N |\hat{\beta}_{\pi_j(i), j} - \hat{\beta}_{m+1,N}(j)|^2 = (N_1 + N_2 - m)|c_{1,m,j}|^2 + N_3|c_{2,m,j}|^2 + 2c_{1,m,j} \sum_{i=m+1}^{N_1+N_2} |u_{\pi_j(i), j} - \bar{u}_{m+1,N}(j)| + 2c_{2,m,j} \sum_{i=m+1}^{N_1+N_2} |u_{\pi_j(i), j} - \bar{u}_{m+1,N}(j)|^2.
\]

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

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

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

\[
\Delta M_{3,j}(m) \equiv M_{3,j}(m) - M_{3,j}(N_1 + N_2) = \frac{N_1 + N_2 - m}{N} |c_{1,m,j}|^2 + \frac{N_3}{N} |c_{2,m,j}|^2 = \frac{N_3 (N_1 + N_2 - m)}{N (N - m)} (\alpha_{1,j}^0 - \alpha_{3,j}^0)^2 \approx \mu_{3mN},
\]

where \( \mu_{3mN} = (N_1 + N_2 - m)/N \). Following the analysis of \( r_{1,j}(m) - r_{1,j}(N_1) \), we can show that \( r_{3,j}(m) - r_{3,j}(N_1 + N_2) = o_p(\mu_{3mN}) \) uniformly in \( m < N_1 + N_2 \). It follows that as \( (N,T) \to \infty \), for any \( j \in S_2 \) we must have

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

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

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

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

\[
P\left( \hat{G}_1(2) = \{\pi_j(1), \ldots, \pi_j(N_1 + N_2)\} \right) \to 1.
\]

Recall here regressor \( j \) is a representative element in set \( S_2 \).

The above proof applies when \( j_1 \in S_2 \). In the second step, our algorithm ensures \( j_2 \in S_1 \) because the segment \( S_{1,N_1+N_2,j} \) contains no break point for any \( j \in S_2 \). Following the analysis in Case 1a, we can
readily show that for any \( j \in S_1 \), we can identify the second break point \( \{N_i\} \) in the second step through our SBSA. As a result, we have \( P(\hat{m}_1 = N_1 + N_2, \hat{m}_2 = N_1) \to 1 \) and \( P(\hat{G}(3) = G^0) \to 1 \) as \( (N,T) \to \infty \). That is, we can identify all three groups w.p.a.1. This completes the proof of the theorem for the case \( K^0 = 3 \). When \( K^0 > 3 \), we need to deal with extra terms. But by similar arguments as that of Bai (1997) and Ke, Li, and Zhang (2016), the proof strategy is essentially the same and the two break points case doesn’t lose generality.

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

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

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

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

\[
IC_1(K) = 2Q_{NT}(\hat{\beta}(K), \hat{\theta}(K)) + pK \cdot \rho_{NT} \geq 2 \min_{1 \leq K < K^0} \min_{\hat{G}(K)} \sigma_0^2(\hat{\beta}(K)) + pK \cdot \rho_{NT} 
\to \sigma_0^2 \text{ as } (N,T) \to \infty.
\]

So we have

\[
P(\hat{K} < K^0) = P(\exists 1 \leq K < K^0, IC_1(K) < IC_1(K^0)) \to 0 \text{ as } (N,T) \to \infty.
\]

Next, we consider the case where \( K^0 < K \leq K^{max} \). By Theorem 3.2, the true group structure will be identified w.p.a.1 when \( K^0 \) is known. When \( K > K^0 \), so we get a further unnecessary refinement of the true group structure. Following the proof of Lemma S1.14 (especially the analysis of \( D_{1NT} \)) in SSP, we can show that

\[
\max_{K^0 < K \leq K^{max}} (\hat{\sigma}_0^2(\hat{\beta}(K^0)) - \hat{\sigma}_0^2(\hat{\beta}(K^{max}))) = O_P(pT^{-1}) \text{ under the condition } p^{3/2}N^{1/2}(\ln N)^9/T \to 0.
\]

It follows that by Assumption A3(iii)

\[
P(\hat{K} > K^0) = P(\exists K^0 < K \leq K^{max}, IC_1(K) < IC_1(K^0))
= P(\exists K^0 < K \leq K^{max}, (T/p)(\hat{\sigma}_0^2(\hat{\beta}(K^0)) - \hat{\sigma}_0^2(\hat{\beta}(K^{max}))) > (K - K^0)T \rho_{NT})
\to 0 \text{ as } (N,T) \to \infty.
\]

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

**Proof of Theorem 3.4:** Let \( D_{NT} = \text{diag}(\sqrt{N_k}T, \ldots, \sqrt{N_k}T, 0, \ldots) \), \( E_{NT} = \{ \hat{G}(K) = G^0 \}, \Xi_{NT} = D_{NT}((\hat{\alpha}^*_1 - \alpha^*_1)\top, \ldots, (\hat{\alpha}^*_K - \alpha^*_K)\top, (\hat{\theta} - \theta^0)\top)\top \), \( \Xi_{NT} = D_{NT}((\hat{\alpha}^*_1 - \alpha^*_1)\top, \ldots, (\hat{\alpha}^*_K - \alpha^*_K)\top, (\hat{\theta} - \theta^0)\top)\top \). Then \( P(E_{NT} \to 1 \text{ as } (N,T) \to \infty \). By Theorems 3.2 and 3.3 and

\[
P(\Xi_{NT} \in C) = P(\Xi_{NT} \in C \text{ and } E_{NT}) + P(\Xi_{NT} \in C \text{ and } E_{NT}^c)
= P(\Xi_{NT} \in C) + o(1),
\]

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

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

\[
\frac{1}{N_kT} \sum_{i=1}^{N_kT} \sum_{\ell=1}^{T} U(w_{it}; \hat{\alpha}_{k_\ell}^*, \hat{\mu}_i; \hat{\alpha}_{k_\ell}^*, \hat{\theta}^*) = 0_{p \times 1} \text{ for } k = 1, \ldots, K^0, \text{ and }
\]

\[
\frac{1}{NT} \sum_{i=1}^{N} \sum_{\ell=1}^{T} W(w_{it}; \hat{\alpha}_{k_\ell}^*, \hat{\mu}_i; \hat{\alpha}_{k_\ell}^*, \hat{\theta}^*) = 0_{q \times 1}.
\]
By Taylor expansions, we have

\[
\begin{bmatrix}
\tilde{H}_{11} & \cdots & \tilde{H}_{10} \\
\vdots & \ddots & \vdots \\
0 & \cdots & \tilde{H}_{K_K0} & \tilde{H}_{K_K0}
\end{bmatrix}
\begin{bmatrix}
\tilde{\alpha}_k^* - \alpha_k^0 \\
\vdots \\
\tilde{\alpha}_k^0 - \alpha_k^0
\end{bmatrix}
= \begin{bmatrix}
\frac{1}{NT} \sum_{i \in G^0_k} \sum_{t=1}^T U(w_{it}; \alpha_k^0, \tilde{\mu}_i(\alpha_k^0, \theta^0), \theta^0) \\
\vdots \\
\frac{1}{N^{K_K0}T} \sum_{i \in G^0_k} \sum_{t=1}^T U(w_{it}; \alpha_k^0, \tilde{\mu}_i(\alpha_k^0, \theta^0), \theta^0)
\end{bmatrix},
\]

where for \( k = 1, \ldots, K^0 \),

\[
\tilde{H}_{kk} = \frac{1}{N_k} \sum_{i \in G^0_k} \tilde{H}_{i,\beta}(\tilde{\alpha}_k, \tilde{\theta}), \\
\tilde{H}_{k\theta} = \frac{1}{N} \sum_{i=1}^N \tilde{H}_{i,\beta}(\tilde{\alpha}_k, \tilde{\theta}), \\
\tilde{H}_{\theta k} = \frac{1}{N} \sum_{i=1}^N \tilde{H}_{i,\beta}(\tilde{\alpha}_k, \tilde{\theta}),
\]

\( \tilde{H}_{i,\beta} \) \((\alpha, \theta)\), \( \tilde{H}_{i,\theta} \) \((\alpha, \theta)\), and \( \tilde{H}_{i,\theta} \) \((\alpha, \theta)\) are defined analogously to \( H_{i,\beta} \) \((\beta, \theta)\), \( H_{i,\beta} \) \((\beta, \theta)\), \( H_{i,\theta} \) \((\beta, \theta)\), and \( H_{i,\theta} \) \((\beta, \theta)\) below \( (3.2) \) with \( \mu_i(\beta, \theta) \) being replaced by \( \tilde{\mu}_i(\alpha, \theta) \), \( \tilde{\alpha}_k \) lies between \( \tilde{\alpha}_k^* \) and \( \tilde{\alpha}_k^0 \) elementwise, and \( \tilde{\theta} \) lies between \( \theta^* \) and \( \theta^0 \). Following the analysis of Theorem 3.1, we can show the L2-consistency of \( \tilde{\alpha}_k^* \) and \( \tilde{\theta}^* \). With this result, we can follow the proof of Lemma S1.13 in SSP (see also the proof of Lemma A.2) and show that

\[
\| \tilde{H}_{kk} - H_{kk} \| = o_P(1), \quad \| \tilde{H}_{k\theta} - H_{k\theta} \| = o_P(1), \quad \| \tilde{H}_{\theta k} - H_{\theta k} \| = o_P(1), \quad \text{and} \quad \| \tilde{H}_{\theta\theta} - H_{\theta\theta} \| = o_P(1),
\]

where, e.g., \( H_{kk} = \frac{1}{N_k} \sum_{i \in G^0_k} E[H_{i,\beta}(\beta^0, \theta^0)], H_{k\theta}, H_{\theta k}, \text{and } H_{\theta\theta} \) are analogously defined.

Let \( S_{kNT} = \frac{1}{\sqrt{NT}} \sum_{i \in G^0_k} \sum_{t=1}^T U(w_{it}; \alpha_k^0, \mu_i(\alpha_k^0, \theta^0), \theta^0) \) for \( k = 1, \ldots, K^0 \) and \( S_{\theta NT} = \frac{1}{\sqrt{NT}} \sum_{i \in G^0_k} \sum_{t=1}^T W(w_{it}; \theta^0, \tilde{\mu}_i(\beta^0, \theta^0), \theta^0) \). As in the proof of Lemma S1.12 in SSP, we apply the second order Taylor expansion to obtain

\[
S_{kNT} = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it} + \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^T U_{it}^\mu[\mu_i(\alpha_k^0, \theta^0)] - \mu_i^0 \]

\[
= S_{kNT, 1} + S_{kNT, 2} + S_{kNT, 3},
\]

\[
S_{\theta NT} = \frac{1}{\sqrt{N T}} \sum_{i=1}^N \sum_{t=1}^T W_{it} + \frac{1}{\sqrt{N T}} \sum_{i=1}^N \sum_{t=1}^T W_{it}^\mu[\mu_i(\beta^0_i, \theta^0)] - \mu_i^0 \]

\[
= S_{\theta NT, 1} + S_{\theta NT, 2} + S_{\theta NT, 3},
\]

where \( [\hat{s_{it,U}}] = [\hat{\mu}_i(\alpha_k^0, \theta^0) - \mu_i^0] U_j^\mu(w_{it}; \alpha_k^0, \mu_i, \theta^0)[\hat{\mu}_i(\alpha_k^0, \theta^0) - \mu_i^0], [\hat{s_{it,W}}] = [\hat{\mu}_i(\beta^0_i, \theta^0) - \mu_i^0] W_j^\mu(w_{it}; \beta^0_i, \tilde{\mu}_i, \theta^0)[\hat{\mu}_i(\beta^0_i, \theta^0) - \mu_i^0] \) denotes the second order partial derivatives of the \( j \)th element of \( U \) \((w_{it}; \alpha_k, \mu_i, \theta)\) with respect to \( \mu_i \), \( W_j^\mu(w_{it}; \beta_i, \tilde{\mu}_i, \theta) \) is similarly defined, and both \( \hat{\mu}_i \) and \( \tilde{\mu}_i \) lie between \( \tilde{\mu}_i(\beta^0_i, \theta^0) \) and \( \mu_i^0 \). To study \( S_{kNT, 3} \) and \( S_{\theta NT, 3} \), we consider the first order Taylor expansion:

\[
0 = \frac{1}{T} \sum_{t=1}^T V(w_{it}; \alpha_k^0, \tilde{\mu}_i(\alpha_k^0, \theta^0))
\]

\[
= \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^T V^\mu(w_{it}; \alpha_k^0, \tilde{\mu}_i(\alpha_k^0, \theta^0), \theta^0) \left[ \hat{\mu}_i(\alpha_k^0, \theta^0) - \mu_i^0 \right],
\]

where \( V(x) \) denotes the first order Taylor expansion of \( V(x) \) around \( x_0 \).
where $\hat{\mu}_i(\alpha_k^0, \theta^0)$ lies between $\hat{\mu}_i(\alpha_k^0, \theta^0)$ and $\mu_i^0$. Solving for $\hat{\mu}_i(\alpha_k^0, \theta^0) - \mu_i^0$ and following the proof of Lemma S1.12 in SSP, we can show that

$$\begin{align*}
[S_{k NT,3}]_j &= \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} T \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right)^T \left( \frac{1}{T} \sum_{t=1}^{T} \sum_{i \in G_k^0} V_{it}^\mu \right)^{-1} W_{it,j}^\mu \left( \frac{1}{T} \sum_{t=1}^{T} V_{it}^\mu \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right) + o_P(1) \\
&= \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it} \right)^T S_{iV}^{-1} S_{iW2,j} S_{iV}^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it} \right) + o_P(1), \\
\text{(B.11)}
\end{align*}$$

$$\begin{align*}
[S_{0 NT,3}]_j &= \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right)^T \left( \frac{1}{T} \sum_{t=1}^{T} \sum_{i \in G_k^0} V_{it}^\mu \right)^{-1} W_{it,j}^\mu \left( \frac{1}{T} \sum_{t=1}^{T} V_{it}^\mu \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right) + o_P(1) \\
&= \frac{1}{2\sqrt{N_k T}} \sum_{i \in G_k^0} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it} \right)^T S_{iV}^{-1} S_{iW2,j} S_{iV}^{-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it} \right) + o_P(1). \\
\text{(B.12)}
\end{align*}$$

To study $S_{k NT,2}$ and $S_{0 NT,2}$, we need to consider the second order Taylor expansion:

$$0 = \frac{1}{T} \sum_{t=1}^{T} V \left( w_{it}; \alpha_k, \hat{\mu}_i(\alpha_k^0, \theta^0), \theta^0 \right)$$

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

where $[\hat{s}_{it,V}]_j = [\hat{\mu}_i(\alpha_k^0, \theta^0) - \mu_i^0] \left[ \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right)^T V_{it}^\mu \left( \frac{1}{T} \sum_{t=1}^{T} \sum_{i \in G_k^0} V_{it}^\mu \right)^{-1} W_{it,j}^\mu \left( \frac{1}{T} \sum_{t=1}^{T} V_{it}^\mu \right)^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right) + o_P(1).$

$$\text{(B.13)}$$

$$\text{where }[\hat{s}_{it,V}]_j = \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right)^T S_{iV}^{-1} V_{it,j} S_{iV}^{-1} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right).$$

With this expression, we can readily show that

$$S_{k NT,2} = -\frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} U_{it}^\mu \left( \frac{1}{T} \sum_{t=1}^{T} V_{it}^\mu \right)^{-1} \left\{ \frac{1}{T} \sum_{t=1}^{T} V_{it} + \frac{1}{2T} \sum_{t=1}^{T} \hat{s}_{it,V} \right\} + o_P(1)$$

$$= -S_{k NT,2} (1) - S_{k NT,2} (2) + o_P(1),$$

where

$$S_{k NT,2} (1) = \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} U_{it}^\mu \left( \frac{1}{T} \sum_{t=1}^{T} V_{it}^\mu \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} V_{it}$$

$$= \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} V_{it} + \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \left[ U_{it}^\mu - \mathbb{E}(U_{it}^\mu) \right] S_{iV}^{-1} \frac{1}{T} \sum_{t=1}^{T} V_{it}$$

$$+ \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} S_{iU} \left[ \frac{1}{T} \sum_{t=1}^{T} V_{it}^\mu \right] - S_{iV}^{-1} \sum_{t=1}^{T} V_{it} + o_P(1)$$

$$= \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} S_{iU} S_{iV}^{-1} V_{it} + \frac{1}{\sqrt{N_k T}} \sum_{i \in G_k^0} \sum_{t=1}^{T} \left[ U_{it}^\mu - \mathbb{E}(U_{it}^\mu) \right] S_{iV}^{-1} V_{it}$$
- \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{s=1}^{T} S_{it} S_{iv}^{-1} (V_{it} - S_{iv}) S_{iv}^{-1} V_{is} + o_p(1)

= \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{s=1}^{T} S_{it} S_{iv}^{-1} V_{it} + \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{s=1}^{T} U_{it} S_{iv}^{-1} V_{is} + o_p(1),

S_{kNT,2}(2) = \frac{1}{2 \sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{t=1}^{T} U_{it} \left( \frac{1}{T} \sum_{t=1}^{T} V_{it} \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} s_{it} V_{is} = \frac{1}{2 \sqrt{N_kT}} \sum_{i \in G_k^0} S_{it} S_{iv}^{-1} R_{iv} + o_p(1),

[R_{iv}]_j = (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})^T S_{iv}^{-1} S_{iw2, j} S_{iv}^{-1} (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it}), and o_p(1) denotes the order of Frobenius norm of the corresponding term here and below (which is made possible under the condition p^{3/2}N^{1/2}(\ln T)/T = o(1) in Assumption A3(i)). It follows that

S_{kNT,2} = \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{t=1}^{T} S_{it} S_{iv}^{-1} V_{it} - \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{s=1}^{T} U_{it} S_{iv}^{-1} V_{is}

- \frac{1}{2 \sqrt{N_kT}} \sum_{i \in G_k^0} S_{it} S_{iv}^{-1} R_{iv} + o_p(1). \quad (B.13)

Similarly, we have

S_{\theta NT,2} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} S_{it} S_{iv}^{-1} V_{it} - \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{s=1}^{T} U_{it} S_{iv}^{-1} V_{is}

- \frac{1}{2 \sqrt{NT}} \sum_{i=1}^{N} S_{iv}^{-1} R_{iv} + o_p(1), \quad (B.14)

where [R_{iv}]_j = (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it})^T S_{iv}^{-1} S_{iw2, j} S_{iv}^{-1} (\frac{1}{\sqrt{T}} \sum_{t=1}^{T} V_{it}). Combining (B.9)–(B.14) yields

S_{kNT} = \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{t=1}^{T} U_{it} - 2 \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} U_{it} - 2 \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{s=1}^{T} U_{it} S_{iv}^{-1} V_{is} + o_p(1) and \ S_{\theta NT} = \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} U_{it} - 2 \frac{1}{\sqrt{NT}} \sum_{i=1}^{N} \sum_{t=1}^{T} U_{it} - 2 \frac{1}{\sqrt{N_kT}} \sum_{i \in G_k^0} \sum_{s=1}^{T} U_{it} S_{iv}^{-1} V_{is} + o_p(1) .

By the Cramér-Wold device and Lindeberg-Feller central limit theorem, we can readily show that S_{\theta NT}^T (S_{\theta NT} + \mathbb{E}_{\theta NT})^T, ..., (S_{\theta NT}^T + \mathbb{E}_{\theta NT}^T)^T is asymptotically normally distributed with mean zero and variance-covariance matrix S^T S. It follows that S_{\theta NT}^T \mathbb{E}_{\theta NT} \rightarrow N (0, S^T \mathbb{E} \mathbb{E}^{-1} S^T) . This completes the proof of the theorem. ■

**Proof of Lemma 4.1**: By the spectral decompositions of A and D_N, we have A = u \Lambda u^T and D_N = U_N \Sigma_N U_N^T = U_{1,N} \Sigma_{1,N} U_{1,N}^T. It follows that for any nonsingular matrix S, we have

D_N = N^{-1} Z_N A Z_N^T = N^{-1} Z_N u \Lambda u^T Z_N^T

= N^{-1} Z_N u S S^{-1} \Lambda (S^{-1})^T u^T Z_N^T = U_{1,N} \Sigma_{1,N} U_{1,N}^T.

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

I_{K^*} = U_{1,N}^T U_{1,N} = N^{-1/2} U_{1,N}^T Z_N u S,
yielding that $S = (N^{-1/2} U_{1,N}^T Z_N u)^{-1}$ provided that $U_{1,N}^T Z_N u$ is nonsingular. By construction, when $S = (N^{-1/2} U_{1,N}^T Z_N u)^{-1}$, $N^{-1}(Z_N u S)^T (Z_N u S) = I_{K^*}$. In addition, we have

$$S^{-1} \Lambda (S^{-1})^T = N^{-1} U_{1,N}^T Z_N (u \Lambda u^T) Z_N^T U_{1,N} = U_{1,N}^T \left( N^{-1} Z_N A Z_N^T \right) U_{1,N},$$

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

To prove (iv), we first show that the rows of $u$ are distinct from each other. Suppose $u$ has two identical rows, which are denoted as row $k$ and row $k'$. We consider rows $k, k'$ and columns $k, k'$ of $A = \alpha^0 \alpha^0^T = u \Lambda u^T$:

$$\begin{bmatrix} c_1 J_2 \cdots c_{K^*} J_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_{K^*} \end{bmatrix} \begin{bmatrix} c_1 J_2 \cdot J_2 \end{bmatrix} = \left( \sum_{k=1}^{K^*} \lambda_k c_k^2 \right) J_2 J_2^T,$$

where $J_2 = (1, 1)^T$ and $c_k$'s are arbitrary scalars as long as $u^T u = I_{K^*}$ is ensured. The last display has identical elements, which implies $\alpha_k^0 \alpha_k^0 = \alpha_k^0 \alpha_k^0 = \alpha_k^0 \alpha_k^0 = \alpha_k^0 \alpha_k^0$. This further implies that

$$\|\alpha_k^0 - \alpha_k^0\|^2 = (\alpha_k^0 - \alpha_k^0)^T (\alpha_k^0 - \alpha_k^0) = 0,$$

i.e., $\alpha_k^0 = \alpha_k^0$ for $k \neq k'$, violating Assumption A2(i). Hence, we can conclude that the rows of $u$ are distinct from each other. Since $S$ is nonsingular, this further ensures that $uS$ has rows that are distinct from each other. Note that if $z_i$ contains 1 in its $k$th element, then $z_i^T u S$ is given by the $k$th row of $u S$. As a result, $z_i^T u S = z_j^T u S$ if and only if $z_i = z_j$ for $i, j = 1, 2, \ldots, N$ because $u S$ has distinct rows.

**Proof of Theorem 4.2:** (i) We first prove that $K_N = K^*$ w.p.a.1. Noting that $\tilde{D}_N - D_N = N^{-1/2} \beta \beta^T - N^{-1/2} \beta \beta^T = N^{-1} (\overline{\beta} - \beta^0) \beta^0 + \beta^0 (\overline{\beta} - \beta^0)^T$,[28] we can readily show that

$$\|\tilde{D}_N - D_N\|^2 = O_P((N-1)\|\overline{\beta} - \beta^0\|^2) = O_P((p/T)). \quad (B.15)$$

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

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

where $\tilde{\mu}_{\ell,N}$ and $\mu_{\ell,N}$ denote the $\ell$th largest eigenvalues of $\tilde{D}_N$ and $D_N$, respectively. Since $D_N$ has rank $K^*$, $\mu_{\ell,N} = 0$ and $\tilde{\mu}_{\ell,N} = O_P((p/T)^{1/2})$ for $\ell \geq K^* + 1$. This implies that $P(K_N > K^*) \rightarrow 0$. By Assumptions A5 and A2 and (4.3), $\tilde{\mu}_{K_N} > \mu_{K^*+1/2} \geq c \min_{1 \leq \ell \leq K^*} \tau_{k}/4 > 0$ w.p.a.1, implying that $P(K_N < K^*) \rightarrow 0$. Consequently, we have $P(K_N = K^*) \rightarrow 1$ as $(N,T) \rightarrow 1$.

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

$$N^{-1/2} \beta^0 = U_N \Sigma_N^{1/2} V_N^T$$

and

$$N^{-1/2} \tilde{\beta} = \tilde{U}_N \tilde{\Sigma}_N^{1/2} \tilde{V}_N^T,$$

where $U_N, \Sigma_N, \tilde{U}_N$ and $\tilde{\Sigma}_N$ are as defined in Section 4.1. $V_N$ is a $p \times p$ matrix such that $V_N^T V_N = I_p$, $\tilde{V}_N$ is a $p \times p$ matrix such that $\tilde{V}_N^T \tilde{V}_N = I_p$. Note that the $(K^* + 1)$th, $\ldots$, $p$th diagonal elements of $\Sigma_N$ are all zero, and the $(K^* + 1)$th, $\ldots$, $p$th diagonal elements of $\tilde{\Sigma}_N$ are all $O_P((p/T)^{1/2})$. We can decompose $\tilde{U}_N,$
\( \hat{\Sigma}_N \), and \( \hat{V}_N \) as follows: \( \hat{U}_N = (\hat{U}_{1,N}, \hat{U}_{2,N}) \), \( \hat{\Sigma}_N = \text{diag}(\hat{\Sigma}_{1,N}, \hat{\Sigma}_{2,N}) \), and \( \hat{V}_N = (\hat{V}_{1,N}, \hat{V}_{2,N}) \), where \( \hat{U}_{1,N} \)

is an \( N \times K^* \) matrix, \( \hat{\Sigma}_{1,N} \) is a \( K^* \times K^* \) diagonal matrix, and \( \hat{V}_{1,N} \) is a \( P \times K^* \) matrix. Analogously, write \( U_N = (U_{1,N}, U_{2,N}) \), \( \Sigma = \text{diag}(\Sigma_{1,N}, \Sigma_{2,N}) \), and \( V_N = (V_{1,N}, V_{2,N}) \), where \( \Sigma_{2,N} \) is a matrix of zeros. Then

\[
N^{-1/2} \beta^0 = U_N \Sigma_{1,N}^{1/2} V_{1,N}^T = U_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^T \tag{B.16}
\]

and

\[
N^{-1/2} \tilde{\beta} = \hat{U}_N \hat{\Sigma}_{1,N}^{1/2} \hat{V}_{1,N} = \hat{U}_{1,N} \hat{\Sigma}_{1,N}^{1/2} \hat{V}_{1,N} + \hat{U}_{2,N} \hat{\Sigma}_{2,N}^{1/2} \hat{V}_{2,N}^T, \tag{B.17}
\]

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

\[
\hat{U}_{1,N}^T U_{1,N} = A_{1N} \Theta_{1N} A_{2N}^T \quad \text{and} \quad \hat{V}_{1,N}^T V_{1,N} = B_{1N} \Theta_{2N} B_{2N}^T,
\]

where \( A_{1N}, A_{2N}, B_{1N} \) and \( B_{2N} \) are all orthogonal matrices, \( \Theta_{1N} = \text{diag}(\cos \theta_1, \ldots, \cos \theta_{K^*}) \), \( \Theta_{2N} = \text{diag}(\cos \theta_{1,2}, \ldots, \cos \theta_{2,K}) \), \( \theta_1, \ldots, \) and \( \theta_{1,K^*} \) are the principal angles between the column spaces of \( \hat{U}_{1,N} \) and \( U_{1,N} \), and \( \theta_{1,2}, \ldots, \) and \( \theta_{2,K} \) are the principal angles between the column spaces of \( \hat{V}_{1,N} \) and \( V_{1,N} \). Let \( O_N = A_{2N} A_{1N} \) and \( \tilde{O} = B_{2N} B_{1N} \). Note that both \( O_N \) and \( \tilde{O} \) are orthogonal matrices. Then by Theorem 4 in Yu, Wang, and Samworth (2015),

\[
\| \hat{U}_{1,N} - U_{1,N} O_N \| = O_P(N^{-1/2}\| \tilde{\beta} - \beta^0 \|) = O_P((p/T)^{1/2}), \tag{B.18}
\]

and

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

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

\[
O_P((p/T)^{1/2}) = N^{-1/2}\| \tilde{\beta} - \beta^0 \| = \| \hat{U}_{1,N} \hat{\Sigma}_{1,N}^{1/2} \hat{V}_{1,N}^T - U_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^T + \hat{U}_{2,N} \hat{\Sigma}_{2,N}^{1/2} \hat{V}_{2,N}^T \|
\]

\[
\geq \| \hat{U}_{1,N} \hat{\Sigma}_{1,N}^{1/2} \hat{V}_{1,N}^T - U_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^T \| - O_P((p/T)^{1/2})
\]

\[
= \| \hat{U}_{1,N} O_N \Sigma_{1,N}^{1/2} \tilde{O} \| \| \hat{V}_{1,N} - U_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^T \| + O_P((p/T)^{1/2})
\]

\[
= \| U_{1,N} O_N \Sigma_{1,N}^{1/2} \tilde{O} \| \| \hat{V}_{1,N} - U_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^T \| + O_P((p/T)^{1/2})
\]

\[
= \| U_{1,N} (O_N \Sigma_{1,N}^{1/2} - \Sigma_{1,N}^{1/2} \tilde{O}) \| \| \hat{V}_{1,N} - U_{1,N} \Sigma_{1,N}^{1/2} V_{1,N}^T \| + O_P((p/T)^{1/2})
\]

\[
= \| O_N \Sigma_{1/2} - \Sigma_{1/2} \tilde{O} \| + O_P((p/T)^{1/2})
\]

\[
= \| O_N \Sigma_{1/2} - \Sigma_{1/2} \tilde{O} \| + O_P((p/T)^{1/2}).
\]

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

Recall that \( \hat{u}_{1,i} = (\hat{u}_{1,i,1}, \hat{u}_{1,i,2}) \), and \( \hat{u}_{2,i} = (u_{1,i}^T, u_{2,i}^T) \) denote the ith row of \( \hat{U}_N = (\hat{U}_{1,N}, \hat{U}_{2,N}) \) and \( U_N = (U_{1,N}, U_{2,N}) \), respectively. Notice that uniformly in \( i \)

\[
N^{-1/2} \hat{\beta}_i^T = \hat{u}_{1,i}^T \hat{\Sigma}_{1,N}^{1/2} \hat{V}_{1,N}^T + \hat{u}_{2,i}^T \hat{\Sigma}_{2,N}^{1/2} \hat{V}_{2,N}^T.
\]

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

\[
N^{-1/2} \hat{\beta}_i^T \hat{V}_{1,N} = \hat{u}_{1,i}^T \hat{\Sigma}_{1,N}^{1/2} \hat{V}_{1,N}.
\]
It follows that \( \max_{1 \leq i \leq N} \| \tilde{u}_{1,i} \| = N^{-1/2} \max_{1 \leq i \leq N} \left\| \Sigma_{1,N}^{-1/2} \hat{\beta}_{i,N} \right\| = N^{-1/2} OP \left( \| \hat{\beta}_i - \beta^0_i \| + \| \beta^0_i \| \right) = O_P (N^{-1/2}) \) by Theorem 3.1 and the additional condition that \( \max_{1 \leq i \leq N} \| \beta^0_i \| = O \left( 1 \right) \) in the theorem. Next, we compare \( \tilde{u}_{1,i} \) and \( u_{1,i} \) through \( \hat{\beta}_i \) and \( \beta^0_i \). By Theorem 3.1 and the fact that \( \hat{\Sigma}_{1,N} - \Sigma_{1,N} = O_P \left( (p/T)^{1/2} \right) \) and that \( \max_{1 \leq i \leq N} \| \tilde{u}_{1,i} \| = O_P (N^{-1/2}) \), we have uniformly in \( i \),

\[
O_P ((p/T)^{1/2} (\ln T)^3) = \| \left( \hat{\beta}_i - \beta_i^0 \right)^\top \tilde{V}_{1,N} = \sqrt{N} \tilde{u}_{1,i} \Sigma_{1,N}^{-1/2} - \sqrt{N} u_{1,i} \Sigma_{1,N}^{-1/2} \tilde{V}_{1,N} \tilde{V}_{1,N} \]

\[
= \sqrt{N} \tilde{u}_{1,i} \Sigma_{1,N}^{-1/2} - \sqrt{N} u_{1,i} \Sigma_{1,N}^{-1/2} \tilde{V}_{1,N} \tilde{V}_{1,N} + \sqrt{N} \tilde{u}_{1,i} \left( \hat{\Sigma}_{1,N} - \Sigma_{1,N} \right)
\]

\[
= \sqrt{N} \left( \tilde{u}_{1,i} - O_N (u_{1,i}) \right)^\top \Sigma_{1,N}^{-1/2} + \sqrt{N} u_{1,i} (O_N \Sigma_{1,N}^{-1/2} - \Sigma_{1,N}^{-1/2} \tilde{V}_{1,N} \tilde{V}_{1,N}) + O_P ((p/T)^{1/2}).
\]

It follows that uniformly in \( i \),

\[
\sqrt{N} \left\| \tilde{u}_{1,i} - O_N (u_{1,i}) \right\| = \left\| - \Sigma_{1,N}^{-1/2} \sqrt{N} (O_N \Sigma_{1,N}^{-1/2} - \Sigma_{1,N}^{-1/2} \tilde{V}_{1,N} \tilde{V}_{1,N}) u_{1,i} + O_P ((p/T)^{1/2} (\ln T)^3) \right\|
\]

\[
\leq \| \Sigma_{1,N}^{-1/2} (O_N \Sigma_{1,N}^{-1/2} - \Sigma_{1,N}^{-1/2} \tilde{O} \tilde{V}_{1,N} \tilde{V}_{1,N}) \| \sqrt{N} u_{1,i} \| + \| \Sigma_{1,N}^{-1/2} [\Sigma_{1,N}^{-1/2} (\tilde{V}_{1,N} \tilde{V}_{1,N} - \tilde{V}_{1,N} \tilde{O} \tilde{V}_{1,N})] \| \sqrt{N} u_{1,i} \| + O_P ((p/T)^{1/2} (\ln T)^3)
\]

\[
\leq \| \Sigma_{1,N}^{-1/2} \| \cdot \| O_N \Sigma_{1,N}^{-1/2} - \Sigma_{1,N}^{-1/2} \tilde{O} \| \cdot \| \sqrt{N} u_{1,i} \| + O_P ((p/T)^{1/2} (\ln T)^3) = O_P (1) O_P (p/T)^{1/2} (\ln T)^3) + O_P (p/T)^{1/2} (\ln T)^3)
\]

This completes the proof of the theorem. \( \blacksquare \)

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

**Proof of Theorem 4.4:** Given the result in Theorem 4.3, the proof of the theorem follows that of Theorem 3.3 and thus omitted. \( \blacksquare \)

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

## C Details of SBSA 2

In this section, we provide the details for SBSA 2.

### Sequential Binary Segmentation Algorithm 2 (SBSA 2)

1. Let \( K \in [1, K_{\text{max}}] \). When \( K = 1 \), there is only one group with the estimate \( \tilde{G}_1 (1) = \{1, \ldots, N\} \).
2. When \( K = 2 \), let \( \tilde{j}_1 = \max_{1 \leq j \leq K_N} \tilde{V}_{1,N} (j) \). Given \( \tilde{j}_1 \), we solve the following minimization problem

\[
\tilde{m}_1 = \arg \min_{1 \leq m < N} \tilde{S}_{1,N} (\tilde{j}_1, m),
\]

which is an estimated break point. Now we have two segments \( \tilde{G}_1 (2) = \tilde{S}_{1,\tilde{m}_1} (\tilde{j}_1) \) and \( \tilde{G}_2 (2) = \tilde{S}_{\tilde{m}_1+1,N} (\tilde{j}_1) \).
3. When $K \geq 3$, we use $\tilde{m}_1, \ldots, \tilde{m}_{K-2}$ denote the break points detected in the previous steps such that $\tilde{m}_1 < \cdots < \tilde{m}_{K-2}$ (perhaps after relabeling) the $K - 2$ break points that have been detected so far. Define

$$
\tilde{j}_{K-1} = \arg \max_{1 \leq j \leq K_N} \sum_{k=1}^{K-1} V_{\tilde{m}_{k-1} + 1, \tilde{m}_k}(j),
$$

$$
\tilde{m}_{K-1}(k) = \arg \min_{\tilde{m}_{k-1} + 1 \leq m < \tilde{m}_k} \hat{S}_{\tilde{m}_{k-1} + 1, \tilde{m}_k}(\tilde{j}_{K-1}, m) \text{ for } k = 1, \ldots, K - 1,
$$

where $\tilde{m}_0 = 0$, $\tilde{m}_{K-1} = N$, and we suppress the dependence of $\tilde{m}_{K-1}(k)$ on $\tilde{j}_{K-1}$. Then $\tilde{m}_{K-1}(k)$ divides $\tilde{G}_k(K - 1)$ into two subsegments, which are labeled as $\tilde{G}_{k1}(K - 1)$ and $\tilde{G}_{k2}(K - 1)$ respectively. Calculate for $k = 1, \ldots, K - 1$,

$$
\hat{S}_{K-1}(k) = \sum_{i \in \tilde{G}_{k1}(K-1)} \left( \hat{U}_{i, \tilde{j}_{K-1}} - \bar{U}_{\tilde{G}_{k1}(K-1)}(\tilde{j}_{K-1}) \right)^2 + \sum_{i \in \tilde{G}_{k2}(K-1)} \left( \hat{U}_{i, \tilde{j}_{K-1}} - \bar{U}_{\tilde{G}_{k2}(K-1)}(\tilde{j}_{K-1}) \right)^2
$$

\[ + \sum_{1 \leq l \leq K-1, l \neq k} \sum_{i \in \tilde{G}_l(K-1)} \left( \hat{U}_{i, \tilde{j}_{K-1}} - \bar{U}_{\tilde{G}_l(K-1)}(\tilde{j}_{K-1}) \right)^2, \]

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

$$
\tilde{k} = \arg \min_{1 \leq \tilde{k} \leq K-1} \hat{S}_{K-1}(k),
$$

which gives 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 $\{\tilde{m}_1, \ldots, \tilde{m}_{K-2}, \tilde{m}_{K-1}(\tilde{k})\}$ and $\{\tilde{G}_1(K), \ldots, \tilde{G}_{\tilde{k}-1}(K - 1), \tilde{G}_{\tilde{k}1}(K - 1), \tilde{G}_{\tilde{k}2}(K - 1), \tilde{G}_{k+1}(K - 1), \ldots, \tilde{G}_{K1}(K)\}$, respectively. Relabel these $K - 1$ break points as $\{\tilde{m}_1, \ldots, \tilde{m}_{K-1}\}$ such that $\tilde{m}_1 < \tilde{m}_2 < \cdots < \tilde{m}_{K-1}$, and the corresponding $K$ groups as $\{\tilde{G}_1(K), \tilde{G}_2(K), \ldots, \tilde{G}_K(K)\}$.

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

D Additional simulation details

Following the suggestion of a referee, we report the goodness-of-fit measure ($R^2$ or pseudo-$R^2$) for each DGP in Table A1. Since $R^2$ is mainly defined for linear models, we consider some pseudo-$R^2$ measures for DGPs 4-5. For DGPs 1-3 which specify linear panel models, we report the value of $R^2$ for the within-group transformed model because the estimation of the parameters of interest is effectively obtained from the within-group transformed model. For DGPs 4-5 that specify two nonlinear models (two-sided censored and one-side censored models), we define the pseudo-$R^2$ as the ratio of the variance of the signal component (e.g., $x_{it}^2 \beta_i + \mu_i$ or $x_{it}^{\text{I}} \beta_i$ in DGP 4) over the variance of the latent variable (e.g., $y_{it}^* = x_{it}^2 \beta_i + \mu_i + \varepsilon_{it}$ in DGP 4). Since $\mu_i$ may or may not be regarded as a part of the signal, we report two pseudo-$R^2$ in this case that are obtained by treating $\mu_i$ as a part of the signal and noise, respectively. Table A1 reports the within-group $R^2$ for DGPs 1-3 and pseudo-$R^2$ for DGPs 4-5 that are obtained as the averages over 500 replications. They range from 0.65 to 0.91.

REFERENCES

Table A1: The $R^2$ or pseudo-$R^2$ for all DGPs

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th>DGP 1</th>
<th>DGP 2</th>
<th>DGP 3</th>
<th>DGP 4 $\mu_i$ as signal</th>
<th>DGP 4 $\mu_i$ as noise</th>
<th>DGP 5 $\mu_i$ as signal</th>
<th>DGP 5 $\mu_i$ as noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10</td>
<td>0.665</td>
<td>-</td>
<td>0.696</td>
<td>0.826</td>
<td>0.652</td>
<td>0.847</td>
<td>0.694</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>0.667</td>
<td>0.900</td>
<td>0.698</td>
<td>0.832</td>
<td>0.663</td>
<td>0.851</td>
<td>0.702</td>
</tr>
<tr>
<td>100</td>
<td>40</td>
<td>0.671</td>
<td>0.912</td>
<td>0.698</td>
<td>0.834</td>
<td>0.656</td>
<td>0.850</td>
<td>0.700</td>
</tr>
<tr>
<td>200</td>
<td>10</td>
<td>0.627</td>
<td>-</td>
<td>0.683</td>
<td>0.831</td>
<td>0.664</td>
<td>0.852</td>
<td>0.704</td>
</tr>
<tr>
<td>200</td>
<td>20</td>
<td>0.652</td>
<td>0.903</td>
<td>0.700</td>
<td>0.833</td>
<td>0.666</td>
<td>0.853</td>
<td>0.706</td>
</tr>
<tr>
<td>200</td>
<td>40</td>
<td>0.678</td>
<td>0.906</td>
<td>0.709</td>
<td>0.833</td>
<td>0.665</td>
<td>0.855</td>
<td>0.710</td>
</tr>
</tbody>
</table>


