11-2016

Homogeneity pursuit in panel data models: Theory and applications

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

Peter C. B. PHILLIPS
Singapore Management University, peterphillips@smu.edu.sg

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

HOMOGENEITY PURSUIT IN PANEL DATA MODELS: THEORY AND APPLICATIONS

By

Wuyi Wang, Peter C. B. Phillips and Liangjun Su

November 2016

COWLES FOUNDATION DISCUSSION PAPER NO. 2063

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Homogeneity Pursuit in Panel Data Models: Theory and Applications*

Wuyi Wang[a], Peter C.B. Phillips[b], Liangjun Su[a]
[a] School of Economics, Singapore Management University
[b] Yale University, University of Auckland, University of Southampton,
& Singapore Management University

November 29, 2016

Abstract

This paper studies estimation of a panel data model with latent structures where individuals can be classified into different groups where slope parameters are homogeneous within the same group but heterogeneous across groups. To identify the unknown group structure of vector parameters, we design an algorithm called Panel-CARDS which is a systematic extension of the CARDS procedure proposed by Ke, Fan, and Wu (2015) in a cross section framework. The extension addresses the problem of comparing vector coefficients in a panel model for homogeneity and introduces a new concept of controlled classification of multidimensional quantities called the segmentation net. We show that the Panel-CARDS method identifies group structure asymptotically and consistently estimates model parameters at the same time. External information on the minimum number of elements within each group is not required but can be used to improve the accuracy of classification and estimation in finite samples. Simulations evaluate performance and corroborate the asymptotic theory in several practical design settings. Two empirical economic applications are considered: one explores the effect of income on democracy by using cross-country data over the period 1961-2000; the other examines the effect of minimum wage legislation on unemployment in 50 states of the United States over the period 1988-2014. Both applications reveal the presence of latent groupings in these panel data.

JEL Classification: C33, C38, C51

Keywords: CARDS; Clustering; Heterogeneous slopes; Income and democracy; Minimum wage and employment; Oracle estimator; Panel structure model

*Correspondence should be addressed to Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; Phone: +65 6828 0386. Su gratefully acknowledges the Singapore Ministry of Education for the Tier-2 Academic Research Fund (AcRF) under grant number MOE2012-T2-2-021 and the funding support provided by the Lee Kong Chian Fund for Excellence. Phillips acknowledges support from the NSF (USA) under Grant SES 12-58258 and Grant NRF-2014S1A2A2027803 from the Korean Government. Email: peter.phillips@yale.edu (P.C.B. Phillips), ljsu@smu.edu.sg (L. Su), wuyi.wang.2013@phdecons.smu.edu.sg (W. Wang).

1 Introduction

Conventional panel data analysis often assumes complete slope homogeneity, which is convenient in practical work and takes full advantage of cross section averaging. However, homogeneity assumptions are frequently rejected in empirical panel studies, as in Hsiao and Tahmiscioglu (1997), Phillips and Sul (2007), Browning and Carro (2007) and Su and Chen (2013). But if complete slope heterogeneity is permitted, estimation can be imprecise or even impractical when the time dimension is very short, thereby losing a key advantage of working with panel data. These considerations motivate the present study and much of the recent research on panel structure modeling.

This paper follows earlier work by Su, Shi, and Phillips (2016, SSP hereafter) by studying a linear panel data model with latent structures that embody unknown homogeneous elements. It is assumed that the cross sectional units can be classified into a small number of groups with homogeneous slopes within each group and heterogeneity across groups. There are many motivating examples for such models in empirical work: in cross country economic growth studies, the presence of possible convergence clubs in the data is often of interest (Phillips and Sul 2007); in financial markets, stock returns in the same sector are commonly thought to share common characteristics (Ke, Fan, and Wu 2015); and in economic geography, location may be a relevant factor in economic performance, leading to spatial geographic groupings in the data (Fan, Lv, and Qi 2011; Bester and Hansen 2016).

The inherent difficulty in studying latent panel structure lies in the unknown nature of the group composition. The practical econometric problem in such cases is that the number of groups is unknown as well as individual group membership within the panel. Since the number of all possible classifications is a Bell number, it is not feasible to try all possible combinations (Shen and Huang 2010). One way to determine the group structure is to use external variables or prior knowledge, such as geographic location and industrial sector composition, to assist in classifying individuals into groups (Bester and Hansen 2016). But this approach is vulnerable to misleading inference when the number of groups or the individual identities are incorrectly specified. Moreover, in many panel data models, there are no natural external variables to assist in classification. Accordingly, much effort has been devoted to determining the unknown panel structure without resorting to the use of external factors. One approach is to use finite mixture models; see Sun (2005), Kasahara and Shimotsu (2009), and Browning and Carro (2010). Another approach adapts the K-means algorithm to panel models in order to form a group structure in the panel; see Lin and Ng (2012), Sarafidis and Weber (2015), Bonhomme and Manresa (2015), and Ando and Bai (2016). In addition, machine learning methods that penalize incorrect choices have also been used to extract group patterns using penalized extremum estimation. In recent work that employs this approach, SSP develop a classification Lasso method (called C-Lasso) in which the penalty takes an additive-multiplicative form that forces the parameters to form into different groups. Coupled with the C-Lasso method, SSP propose BIC-type information criteria to determine the number of groups. In additional work, Lu and Su (2016) propose a direct testing procedure to identify the

group number in this linear panel structure model.

When a panel data model has a latent group structure, the problem falls within the framework of high dimensional modeling with parameters that may lie in a low dimension subspace. This type of regression model is now a major research area in statistics; see, for example, the monograph by Bühlmann and van der Geer (2011). Since the work of Tibshirani (1996) and Fan and Li (2001), much of the statistical research has concentrated on sparsity, where a large dimensional space is simplified by zeroing out many elements to reduce dimension. Sparsity may be regarded as a special case of homogeneity where the commonality arises from a shared zero coefficient value. Much effort has been devoted to the study of homogeneity in parameters. When there is a natural variable to define neighborhood, the idea of fused lasso (Tibshirani et al. 2005) can be used to study homogeneity. When there is no such natural variable, exhaustive pairwise penalties have been proposed to address homogeneity. For instance, Bondell and Reich (2008) design a method called OSCAR (octagonal shrinkage and clustering algorithm for regression) where the octagonal penalty is imposed on all pairs of coordinates to form clusters; and Shen and Huang (2010) propose to use a truncated L_1 penalty on all pairs of predicators to extract a grouping structure.

Ke, Fan, and Wu (2015, KFW hereafter) explore homogeneity in regressions by designing a method called CARDS (clustering algorithm in regression via data-driven segmentation). They first estimate the parameters by OLS to obtain preliminary estimates. Then the fitted coefficients are ranked from smallest to largest and ordered partition sets (groups) of regressors are constructed based on this ranking. Penalized least squares (PLS) regressions are run to obtain the final estimates where the penalties are imposed on both the within group coefficient differences and neighboring group coefficient differences. KFW show that CARDS can produce oracle estimates with probability approaching 1 (w.p.a.1).[1] They remark that CARDS can be extended to panel data models, but their simple extension does not explore the panel data structure fully and there are conceptual and technical complications that prevent immediate implementation.

This paper extends the CARDS method to panel structure models in a systematic way that deals with these complications. The new method is called Panel-CARDS and it differs from CARDS in two ways. First, Panel-CARDS imposes penalties on slope vector differences while CARDS does so on individual slope differences. In a panel data model with $p > 1$ regressors, the KFW CARDS method treats each of the p regressors as an independent unit, constructs the penalty term for each regressor as in the cross section framework, and then adds all p penalty terms to the least squares objective function to form the PLS extremum estimation problem. Usually, different regressors will report different classification results which the new Panel-CARDS can avoid. Second, to use more information from the preliminary estimates, we extend the ordered segmentation concept proposed in KFW to the segmentation net, which enables us to extract groups more accurately. Just as CARDS for cross section data or the SSP C-Lasso for panel data, Panel-CARDS can identify the number of groups and estimate the parameters at the same time.

In addition, we relax various conditions used in KFW and SSP. For example, KFW require non-

[1] An oracle estimate is one that one can achieve by knowing the exact group structure.

stochastic regressors and sub-Gaussian errors whereas we permit random regressors, include lagged dependent variables, and replace sub-Gaussian requirements by moment conditions. Further, SSP require the number of elements in each group to be divergent with sample size and the number of groups to be fixed, whereas we allow the number of elements in each group and the number of groups to be either fixed or divergent to infinity.

We provide two empirical applications of this new panel classification procedure. The first application re-investigates relationships between income and democracy, a matter that has attracted considerable interest among political economists (c.f. Acemoglu et al. 2008). In different countries, the effect of income on democracy might be similar or might differ. Our methods reveal a positive relationship between the two variables in some countries (e.g., South Korea, Japan, Romania, and Spain), a negative relationship between them in other countries (e.g., Iran and Malaysia), and little evidence of a relationship between income and democracy in the remainder (e.g., China and Singapore). In particular, the democracy indices for the countries in the last group have not changed much over the last four decades despite their rapid economic growth. For this reason, estimation and inference based on a fully homogeneous panel data model might well lead to misleading inferences about a generic form of this relationship. Our approach allows for a panel structure of possibly homogeneous and heterogeneous effects of income on democracy. The empirical implementation of Panel-CARDS estimation with these data identifies three latent groupings among the 74 countries corresponding to positive, negative, and indifferent associations between income and democracy.

Our second application studies the impact of minimum wage legislation on unemployment in the United States. This topic has been widely studied in labor economics but has generated some controversy over the last two decades with different research drawing different conclusions (c.f. Dube et al. 2010). This divergence in past empirical research motivates the use of a more flexible modeling framework in which latent panel structures allow for unobserved slope heterogeneity across groups. Panel-CARDS estimation identifies two groupings of states. In one group, a rise in the minimum wage is associated with a decrease in the unemployment rate whereas the opposite effect is observed in the other group. One notable finding from our study is that the two groups have a surprisingly regular geographic distribution on the map, in which the top 15 largest states in terms of GDP all lie in the same group despite the fact that no geographic or economy scale information is used in the Panel-CARDS. This finding indicates that the data-based methodology of Panel-CARDS can help in the discovery of relevant geographic determinants.

The rest of the paper is organized as follows. Section 2 introduces the panel structure model and the Panel-CARDS algorithm. Section 3 develops the properties and asymptotic theory of Panel-CARDS. Simulation performance in finite samples is studied in Section 4. Section 5 applies the methodology to study the effect of income on democracy and that of the minimum wage on unemployment. Section 6 concludes. Proofs are given in the Appendix.

Notation. For integer n, $\mathbb{R}^n$ denotes n dimensional Euclidean space. For vector $\boldsymbol{\alpha} \in \mathbb{R}^n$, the L_q norm of $\boldsymbol{\alpha}$ is defined as $\|\boldsymbol{\alpha}\|_q = (\sum_{j=1}^n |\alpha_j|^q)^{1/q}$ with $1 \leq q < \infty$. When $q = 2$, we abbreviate

$\|\cdot\|_2$ as $\|\cdot\|$. Let $\|\boldsymbol{\alpha}\|_\infty = \max_{1\leq j\leq n}|\alpha_j|$. For a square matrix A of order n, its induced L_q norm is $\|A\|_q = \max_{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_q=1}\|A\boldsymbol{\alpha}\|_q$. When $q=2$, we omit the subscript q. When A is symmetric, we denote by $\mu_{\max}(A)$ and $\mu_{\min}(A)$ the largest and smallest eigenvalues of A. For two real numbers a and b, $a\vee b$ denotes $\max(a,b)$. For two real sequences $\{a_k\}$ and $\{b_k\}$, $a_k \gg b_k$ means that $a_k/b_k \to \infty$ as $k\to\infty$.

2 Panel-CARDS

This section introduces the panel structure model and reviews the original CARDS procedure before developing the Panel-CARDS algorithm.

2.1 Panel structure models

Following SSP, we consider a panel data model with latent group structure

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta}_i^0 + \mu_i + \varepsilon_{it},\ i=1,...,N,\ t=1,...,T, \tag{2.1}$$

where $\mathbf{x}_{it} = (x_{it1},...,x_{itp})'$ is a $p\times 1$ vector of regressors, μ_i is the individual fixed effect which may be correlated with $\mathbf{x}_{it}$, ε_{it} is an idiosyncratic error term with zero mean, and $\boldsymbol{\beta}_i^0$ is a $p\times 1$ vector of slope parameters that admit a possible grouping structure of the form

$$\boldsymbol{\beta}_i^0 = \begin{cases} \boldsymbol{\alpha}_1^0 & \text{if } i\in G_1^0 \\ \vdots & \vdots \\ \boldsymbol{\alpha}_K^0 & \text{if } i\in G_K^0 \end{cases}. \tag{2.2}$$

Here $\boldsymbol{\alpha}_l^0 \neq \boldsymbol{\alpha}_k^0$ for any $l\neq k$, and $\mathcal{G} = \{G_1^0, G_2^0, ..., G_K^0\}$ forms a partition of $\{1,2,...,N\}$. Let $N_k = \left|G_k^0\right|$ denote the cardinality of G_k^0, $k=1,...,K$. Let

$$\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_1', ..., \boldsymbol{\alpha}_K')' \text{ and } \boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_N')'. \tag{2.3}$$

The true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are denoted by $\boldsymbol{\alpha}^0$ and $\boldsymbol{\beta}^0$. We intend to apply a CARDS-type approach to identify the group structure $\mathcal{G}$ and to estimate the group-specific regression coefficients $\boldsymbol{\alpha}^0$ simultaneously.

2.2 The original CARDS

KFW consider the cross sectional linear regression model

$$y_i = \mathbf{x}_i'\mathbf{b}^0 + e_i,\quad i=1,...,n, \tag{2.4}$$

where $\mathbf{x}_i$ is a $p\times 1$ vector of regressors, the e_i's are independently and identically distributed (i.i.d.) error terms with mean zero and variance σ^2, and $\mathbf{b}^0$ is a $p\times 1$ vector of parameters of interest.

They assume that there is a partition $\mathcal{H} = \{H_1^0, H_2^0, ..., H_K^0\}$ of the parameter indices $\{1, 2, ..., p\}$ such that

$$b_\iota^0 = \begin{cases} a_1^0 & \text{if } \iota \in H_1^0 \\ \vdots & \vdots \\ a_K^0 & \text{if } \iota \in H_K^0 \end{cases}, \tag{2.5}$$

where a_k^0 is the common parameter value shared by all members in H_k^0, and $a_l^0 \neq a_k^0$ for any $l \neq k$.

Note that in (2.2), cross sectional individuals have the grouping structures and the $\boldsymbol{\beta}_i$'s are vectors. While in (2.5), regressors have the group structure and the b_ι's are scalars. This is a fundamental difference in the two models that is due to the structure of cross sectional and panel data. Without loss of generality, we assume $a_1^0 < a_2^0 < \cdots < a_K^0$.

The basic idea in the KFW CARDS algorithm is to use preliminary estimates to construct a ranking of the estimates that leads to an ordered segmentation. The formal definition of ordered segmentation is as follows.

Definition 1. *For a segmentation $\mathcal{B} = \{B_1, ..., B_L\}$ which is a partition of the set $\{1, ..., p\}$, $\mathcal{B}$ is called an ordered segmentation if $\max_{\iota \in B_l} b_\iota^0 \leq \min_{\iota \in B_{l+1}} b_\iota^0$ for $l = 1, ..., L-1$.*[2]

Once an ordered segmentation is determined, penalized least squares (PLS) can be used to extract potential groupings of the regressors. This is performed in the following steps:

- **Preliminary Estimation**: Obtain a consistent preliminary estimate $\tilde{\mathbf{b}}$ of $\mathbf{b}$. For model (2.4) with $p \ll n$, we can use the OLS estimate as the preliminary estimate.

- **Preordering**: Sort the coefficients in $\tilde{\mathbf{b}}$ in ascending order. The rank mapping $\tau(\cdot)$ is determined by the ranking relation below

  $$\tilde{b}_{\tau(1)} \leq \tilde{b}_{\tau(2)} \leq \cdots \leq \tilde{b}_{\tau(p)},$$

  where $\tilde{b}_{\tau(\iota)}$ is the ι-th smallest value in $\{\tilde{b}_l : 1 \leq l \leq p\}$.

- **Ordered Segmentation**: Let $\delta > 0$ be a tuning parameter. Find all the indexes $i_2 < i_3 < \cdots < i_L$ such that the gaps

  $$|\tilde{b}_{\tau(\iota)} - \tilde{b}_{\tau(\iota-1)}| > \delta, \qquad \iota = i_2, ..., i_L.$$

  Construct the ordered segmentation $\mathcal{B}$ as

  $$B_l = \{\tau(i_l), \tau(i_l + 1), ..., \tau(i_{l+1} - 1)\}, \qquad l = 1, ..., L,$$

  where $i_1 = 1$ and $i_{L+1} = p + 1$.

[2] In brief, an ordered segmentation in KFW means that the order of $\mathbf{b}^0$ is preserved.

- **CARDS Penalty Function**: Next construct a penalty function with two parts. One is the within-segmentation penalty and the other is a penalty between neighboring segmentations. The penalty function $\rho_\lambda(\cdot)$ used here is the smoothly clipped absolute deviation (SCAD) penalty of Fan and Li (2001). The within-segment penalty drives parameters in the same segment to converge to each other when they are actually in the same true group. The other penalty term penalizes neighboring segment pairs. If the preliminary estimates are accurate enough, the neighboring pairs may be true neighbors or in the same group. In both cases, the SCAD penalty function can help achieve homogeneous values for parameters in the same group and heterogeneous values across groups. The form of the CARDS penalty is given by the expression

$$P_{\mathcal{B},\lambda_1,\lambda_2}(\mathbf{b}) = \underbrace{\sum_{l=1}^{L-1} \sum_{\iota\in B_l,\kappa\in B_{l+1}} \rho_{\lambda_1}(|b_\iota - b_\kappa|)}_{\text{between-segment penalty}} + \underbrace{\sum_{l=1}^{L} \sum_{\iota\in B_l,\kappa\in B_l} \rho_{\lambda_2}(|b_\iota - b_\kappa|)}_{\text{within-segment penalty}}. \tag{2.6}$$

- **Penalized Least Squares**: Solve the PLS problem

$$Q_n(\mathbf{b}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b})^2 + P_{\mathcal{B},\lambda_1,\lambda_2}(\mathbf{b}). \tag{2.7}$$

  Given the tuning parameter vector $\boldsymbol{\lambda} \equiv (\delta, \lambda_1, \lambda_2)'$, we obtain an estimate $\hat{\mathbf{b}}(\boldsymbol{\lambda})$ which may be used to obtain the estimated number of groups, $K(\boldsymbol{\lambda})$. Let $\sigma_n^2(\boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}[y_i - \mathbf{x}_i'\hat{\mathbf{b}}(\boldsymbol{\lambda})]^2$.

- **Select Tuning Parameters by BIC**: Choose $\boldsymbol{\lambda}$ to minimize

$$\text{BIC}(\boldsymbol{\lambda}) = \ln\left(\sigma_n^2(\boldsymbol{\lambda})\right) + K(\boldsymbol{\lambda})\frac{\ln n}{n}. \tag{2.8}$$

The CARDS method has a straightforward extension to panel data models. In KFW's Experiment 5, which is a panel structure model as described above, they construct the CARDS penalty for each regressor and then add them up together. So the penalized objective function can be expressed as

$$Q_{NT}(\boldsymbol{\beta}) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta}_i)^2 + \sum_{\iota=1}^{p} P_{\mathcal{B}_\iota,\lambda_{1\iota},\lambda_{2\iota}}\left(\underline{\boldsymbol{\beta}}_\iota\right), \tag{2.9}$$

where $\underline{\boldsymbol{\beta}}_\iota = (\beta_{1\iota}, ..., \beta_{N\iota})'$ collects the coefficients of the ι-th regressors for all N cross sectional units. This method works but has two serious drawbacks. First, it involves $3p$ tuning parameters, which is excessive for a Lasso procedure when $p \geq 2$. Second, since $P_{\mathcal{B}_\iota,\lambda_{1\iota},\lambda_{2\iota}}(\underline{\boldsymbol{\beta}}_\iota)$ imposes a penalty that is specific to regressor ι only, the classification errors tend to accumulate through the addition of the p sets of penalty terms. Below, we introduce the modified procedure Panel-CARDS which removes these drawbacks and provides an improvement over the basic CARDS procedure for panel data applications.

2.3 Rank mapping in the panel data model

Without the latent group structure (2.2), we can estimate the model (2.1) directly. After concentrating out the fixed effects, we obtain the objective function

$$L_{NT}(\boldsymbol{\beta}) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i)^2, \tag{2.10}$$

where $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - \bar{\mathbf{x}}_i$ and $\tilde{y}_{it} = y_{it} - \bar{y}_i$ with $\bar{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}$ and $\bar{y}_i = \frac{1}{T}\sum_{t=1}^{T}y_{it}$. Solving the optimization problem yields the OLS estimates $\tilde{\boldsymbol{\beta}}_i = (\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}')^{-1}(\frac{1}{T}\sum_{t=1}^{T}\tilde{\mathbf{x}}_{it}\tilde{y}_{it})$ for $i = 1, 2, ..., N$.

Define $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \tilde{\boldsymbol{\beta}}_2', ..., \tilde{\boldsymbol{\beta}}_N')'$, a $pN \times 1$ vector, and $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \cdots, \tilde{\boldsymbol{\beta}}_N)$, a $p \times N$ matrix. To use CARDS, we need to have a rank mapping over the cross section dimensions according to the vector $\tilde{\boldsymbol{\beta}}$. If $p = 1$, the problem is exactly the same as the cross sectional case. We just sort elements in $\tilde{\boldsymbol{\beta}}$ in ascending order. But usually $p > 1$, and we face the awkward problem of ranking N column vectors in $\tilde{\mathbf{B}}$, which is not trivial. Reasonable ranking rules should satisfy the following set of conditions:

1. *Unrestricted Domain*: All $N!$ kinds of rankings are possible.
2. *Unanimity*: If all p elements in $\tilde{\boldsymbol{\beta}}_i$ are less than the corresponding elements in $\tilde{\boldsymbol{\beta}}_l$, then $\tilde{\boldsymbol{\beta}}_i$ should rank before $\tilde{\boldsymbol{\beta}}_l$.
3. *Independence of Irrelevant Alternatives*: The rankings of $\tilde{\boldsymbol{\beta}}_i$ and $\tilde{\boldsymbol{\beta}}_l$ are not affected by $\tilde{\boldsymbol{\beta}}_k$ where $k \neq i$ and $k \neq l$. Otherwise, the ranking result might be totally changed by the introduction of a new individual $N + 1$.

The above three criteria connect the problem of ranking vectors with a famous impossibility theorem in social choice theory. In that setting, we take $\iota = 1, 2, ..., p$ as voters (each row of $\tilde{\mathbf{B}}$) and the numeric ranking as a preference order. According to Arrow's impossibility theorem (e.g., Mas-Colell et al. 1995, p.796), to satisfy all the above three criteria we will inevitably end up with a "dictator", which means our ranking must be totally determined by a single "voter". So we have the following theorem.

Theorem 2.1 *To satisfy the unrestricted domain, unanimity, and independence of irrelevant alternatives assumptions, the rankings of N preliminary vector estimates (columns of matrix $\tilde{\mathbf{B}}$) must be totally determined by the rankings of the preliminary estimates of the coefficients of one regressor, i.e., one particular row of $\tilde{\mathbf{B}}$.*

Now we only need to select a proper element ι^* from $\{1, 2, ..., p\}$ as the "dictator". Noting that we want to obtain the heterogeneity/homogeneity information from preliminary estimates across individuals, it is wise to choose the regressor whose slope coefficient estimates have larger variation across individuals than the others. Let ι^* denote the index of the regressor which has the largest

variation across individuals for its coefficient estimates. Then we can sort $\{\tilde{\beta}_{i\iota^*},\, i = 1, 2, ..., N\}$ to obtain the order

$$\tilde{\beta}_{\tau(1)\iota^*} \leq \tilde{\beta}_{\tau(2)\iota^*} \leq \cdots \leq \tilde{\beta}_{\tau(N)\iota^*}. \tag{2.11}$$

To proceed, we need to define an admissible segmentation.

Definition 2. *For a segmentation $\mathcal{B} = \{B_1, \cdots, B_L\}$ of the set $\{1, ..., N\}$ with true grouping structure $\mathcal{G} = \{G_1^0, G_2^0, \cdots, G_K^0\}$, let $V_{kl} = G_k^0 \cap B_l$ if we have: (i) for each k, G_k^0 is properly segmented by $\mathcal{B}$—there exist d_k and u_k such that $d_k \leq u_k$, $G_k^0 = \cup_{l=d_k}^{u_k} V_{kl}$, and $V_{kl} = B_l$ for $d_k < l < u_k$; (ii) for each l, there exist a_l and b_l such that $a_l \leq b_l$, $B_l = \cup_{k=a_l}^{b_l} V_{kl}$, and $V_{kl} = G_k^0$ for $a_l < k < b_l$, then the segmentation $\mathcal{B}$ is called an admissible segmentation.*

Note that when $p = 1$, an ordered segmentation is also an admissible segmentation. Intuitively, the admissible segmentation $\mathcal{B}$ should segment the individuals in a way that no true group members of G_k^0 fall to disconnected B_l's. Consider a simple illustrative example where $N = 10$ and $\mathcal{G} = \{G_1^0, G_2^0, G_3^0\}$ with $G_1^0 = \{1, 2, 3\}$, $G_2^0 = \{4, 5, 6\}$ and $G_3^0 = \{7, 8, 9, 10\}$. If from (2.11) together with a tuning parameter δ we have a segmentation comprised of $B_1 = \{2, 3\}$, $B_2 = \{1, 5\}$, $B_3 = \{4, 6, 7\}$, $B_4 = \{9, 10\}$, and $B_5 = \{8\}$, then we can easily verify that the segmentation is admissible by Definition 2.[3] But the segmentation $\mathcal{B} = \{B_1, \cdots, B_5\}$ with $B_1 = \{2, 3\}$, $B_2 = \{1, 5, 7\}$, $B_3 = \{4, 6\}$, $B_4 = \{9, 10\}$ and $B_5 = \{8\}$ is not admissible.

To rank vectors, we need to make sure the admissibility of a segmentation. But the last requirement is not always ensured and it may be difficult to satisfy when the true group-specific coefficients exhibit some patterns. To see this, suppose $p = 2$ in the above example and the true group-specific coefficients are given by

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left(\begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1.5 \end{bmatrix} \right). \tag{2.12}$$

If we choose $\iota^* = 1$, say, then there is no chance to obtain an admissible segmentation no matter how accurate the preliminary estimates are. On the other hand, if we will choose $\iota^* = 2$, then it is not hard to obtain an admissible segmentation asymptotically provided that the preliminary estimates are consistent. If, for the above example, $p = 3$ and the true group-specific parameter values are given by

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left(\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \right), \tag{2.13}$$

then it is generally impossible to obtain an admissible segmentation no matter which regressor is chosen to construct the ranking and whether the preliminary estimates are consistent or not. The latter case needs special attention and will be addressed in Section 2.5 below.

[3] One possible ranking is: $\tilde{\beta}_{2\iota^*} \leq \tilde{\beta}_{3\iota^*} \leq \tilde{\beta}_{1\iota^*} \leq \cdots \leq \tilde{\beta}_{9\iota^*} \leq \tilde{\beta}_{10\iota^*} \leq \tilde{\beta}_{8\iota^*}$, with $\tilde{\beta}_{1\iota^*} - \tilde{\beta}_{3\iota^*} > \delta, \cdots \tilde{\beta}_{8\iota^*} - \tilde{\beta}_{10\iota^*} > \delta$. Besides, $V_{11} = \{2, 3\}$, $V_{12} = \{1\}$; $V_{22} = \{5\}$, $V_{23} = \{4, 6\}$; $V_{33} = \{7\}$, $V_{34} = \{9, 10\}$, $V_{35} = \{8\}$.

2.4 Construction of the basic Panel-CARDS

Now suppose we have an admissible segmentation $\mathcal{B} = \{B_1, B_2, ..., B_L\}$. As in the KFW CARDS algorithm, we propose the following hybrid penalty

$$P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \underbrace{\sum_{l=1}^{L-1} \sum_{i\in B_l, j\in B_{l+1}} p_{\lambda_1}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1)}_{\text{between-segment penalty}} + \underbrace{\sum_{l=1}^{L} \sum_{i\in B_l, j\in B_l} p_{\lambda_2}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1)}_{\text{within-segment penalty}}, \tag{2.14}$$

where $p_\lambda(\cdot)$ is the SCAD function of Fan and Li (2001). Here we use L_1 distance to measure the difference between coefficient pairs. For L_q distance, the larger q is, the more weight is placed on the large elements in the norm. In the extreme where $q = \infty$, only the largest element in the vector matters. By adding the penalty term (2.14) to the original objective function (2.10), we obtain the following PLS objective function

$$Q_{NT}(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{2.15}$$

We call the above procedure basic Panel-CARDS. For implementation, we may apply the local linear approximation (LLA) algorithm to obtain the solution. We start from the initial solution and update it by solving the following iterative minimization problem

$$\hat{\boldsymbol{\beta}}^{(s+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ L_{NT}(\boldsymbol{\beta}) + R(\hat{\boldsymbol{\beta}}^{(s)}; \boldsymbol{\beta}) \right\}, \tag{2.16}$$

where $R(\hat{\boldsymbol{\beta}}^{(s)}; \boldsymbol{\beta}) = \sum_{l=1}^{L-1} \sum_{i\in B_l, j\in B_{l+1}} p'_{\lambda_1}(\|\hat{\boldsymbol{\beta}}_i^{(s)} - \hat{\boldsymbol{\beta}}_j^{(s)}\|_1)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1 + \sum_{l=1}^{L} \sum_{i\in B_l, j\in B_l} p'_{\lambda_2}(\|\hat{\boldsymbol{\beta}}_i^{(s)} - \hat{\boldsymbol{\beta}}_j^{(s)}\|_1)\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1$. Noting that the objective function in (2.16) is convex, we can apply a standard convex optimization package to obtain the solution. We use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ to denote the final solution.

Evidently, the performance of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ depends on the choice of $\boldsymbol{\lambda}$. Following SSP, we can choose $\boldsymbol{\lambda} = (\delta, \lambda_1, \lambda_2)'$ to minimize the following information criterion

$$\text{IC}(\boldsymbol{\lambda}) = \ln\left(\sigma^2_{NT}(\boldsymbol{\lambda})\right) + pK(\boldsymbol{\lambda})\frac{1}{2\sqrt{NT}}, \tag{2.17}$$

where $\sigma^2_{NT}(\boldsymbol{\lambda})$ and $K(\boldsymbol{\lambda})$ are estimates of σ^2 and number of groups associated with $\boldsymbol{\lambda}$.[4]

2.5 Construction of the advanced Panel-CARDS

In these last two subsections, we study the admissible segmentation and then construct PLS estimates based upon it. This is a direct extension of CARDS from the cross sectional case to panel data. Nevertheless, such a method does not work in some sparse cases. For example, for the

[4] Note that the value of δ determines the number of segments L in $\mathcal{B}$. Too small or too large a δ will generate too many or too few segments which are not ideal in achieving correct identification. In practice, we find it is helpful to set the number of segments directly, which is also easy to control. For example, when $N = 100$, we try $L = 10$, 20, and 30. The choices of λ_1 and λ_2 depend on the value of coefficients we use in the DGP. Generally speaking, when the coefficients are large, the tuning parameters λ_1 and λ_2 are large correspondingly.

group-specific parameters considered in (2.13), whichever regressor is chosen, we cannot obtain an admissible segmentation no matter how accurate the preliminary estimates are. So the basic Panel-CARDS method fails to work asymptotically in this case and we need to consider an alternative way to obtain robust classification and estimation.

In the example introduced at the end of section 2.3, we can only extract partial information about the grouping property from any single regressor. Naturally, we want to combine information from all regressors in a proper way to derive the true grouping property. Based on this idea, we propose an advanced version of Panel-CARDS which can be regarded as an extension of the basic Panel-CARDS procedure.

In the basic Panel-CARDS method, the admissible segmentation is used to construct both the within segment penalty and the neighboring segments penalty. Compared with the number of exhaustive pairwise penalty terms, the number of penalty terms in basic Panel-CARDS is much smaller. This tends to eliminate penalty terms that are necessary in recovering the true grouping properties when the segmentation is not admissible. In practice, it is desirable to maintain a balance between keeping the number of penalty terms small and having enough penalty terms to extract the grouping structure.

Definition 3. *Let $\mathcal{G} = \{G_1^0, G_2^0, ..., G_K^0\}$ denote the true grouping structure. Given R segmentations $\mathcal{B}_{\iota_1}, ..., \mathcal{B}_{\iota_R}$, if for any G_k^0, there exists a $\mathcal{B}_{\iota_r}$ such that G_k^0 can be properly segmented by $\mathcal{B}_{\iota_r}$ as defined in Definition 2, then $\mathcal{N} \equiv \{\mathcal{B}_{\iota_1}, ..., \mathcal{B}_{\iota_R}\}$ is called an admissible segmentation net.*

Given an admissible segmentation net $\mathcal{N} = \{\mathcal{B}_{\iota_1}, \cdots, \mathcal{B}_{\iota_R}\}$, the advanced Panel-CARDS algorithm is as follows:

- For each $\mathcal{B}_{\iota_r}$, we construct the penalty function $P_{\mathcal{B}_{\iota_r},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ as introduced in (2.14).
- For the admissible segmentation net $\mathcal{N}$, the total penalty is

$$P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \sum_{r=1}^{R} P_{\mathcal{B}_{\iota_r},\lambda_1,\lambda_2}(\boldsymbol{\beta}).$$

- We choose $\boldsymbol{\beta}$ to minimize the following PLS function:

$$Q_{NT}^*(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{2.18}$$

Advanced Panel-CARDS reduces to basic Panel-CARDS in case $R = 1$. When $R > 1$, $P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ contains all the penalty terms that are necessary to recover the true grouping structure. The basic idea of an admissible segmentation net is to extract an adequate amount of information from the preliminary estimates: not too much because we don't use exhaustive pairwise penalties which are challenging in computation and not accurate in statistical inference (as in KFW); and not too few, in order to handle the sparse parameters case introduced at the end of Section 2.3.[5]

[5] Its existence follows directly from Theorem 3 of KFW.

Although here we need the admissible segmentation net to properly segment every true group, we show in DGP 1 below through simulations that when this condition is mildly violated (e.g., there exists one group which cannot be properly segmented by any segmentation), the classification based on the basic Panel-CARDS may still perform well in finite samples.

2.6 Hierarchical clustering

When the signal noise ratio is small or the time period T is relatively small, the preliminary estimates might be quite different from the true parameter values. In such cases, both the basic and advanced Panel-CARDS procedures may produce an estimated number of groups that is greater than the true number of groups, and some estimated groups may only contain few individuals. It is hard, if it is possible at all, to disentangle whether such small groups are the correct groups or are generated because of mis-classification. However, if we have some *a priori* knowledge about the grouping structure, we can use this knowledge during the Panel-CARDS implementation. Following the idea presented in Park et al. (2007), we can use hierarchical clustering to combine members in small groups into large groups. For example, if we know each group contains more than $\eta = 2\%$ of individuals, then we can easily incorporate such information in the procedure. The details will be introduced in the simulation section.

3 Asymptotic Analysis of Panel-CARDS

This section analyzes the large sample properties of the Panel-CARDS algorithm.

3.1 Assumptions

To proceed, we define some notation. Let $\tilde{\mathbf{x}}_i = (\tilde{\mathbf{x}}_{i1}, \cdots, \tilde{\mathbf{x}}_{iT})'$, $\tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, ..., \tilde{y}_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}, ..., \mathbf{x}_{iT})'$, and $\mathbf{y}_i = (y_{i1}, ..., y_{iT})'$. Let $\max_{i,t}$ denote $\max_{1\leq i\leq N} \max_{1\leq t\leq T}$. Let $\rho_j(s) = \lambda_j^{-1} p_{\lambda_j}(s)$ and $\bar{\rho}_j(s) = \rho_j'(s) = p_{\lambda_j}'(|s|)\text{sgn}(s)$ where $p_{\lambda_j}'(s) = dp_{\lambda_j}(s)/ds$ for $j = 1, 2$. Let $b_{NT} = \frac{1}{2}\min_{1\leq k<j\leq K} \|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_j^0\|_1$. Given $\{G_k^0\}$ and segmentation $\{B_1, \cdots, B_L\}$, we define $\phi_k = N_k / \min\{N_k^3, \min_{d_k\leq l\leq u_k} |B_l|^2\}$. Note that $1/N_k^2 \leq \phi_k \leq N_k$. We use $(N_k, T) \to \infty$ to signify that N_k and T pass to infinity jointly.

We make the following assumptions.

Assumption A1.(i) *For each i, $\{(\mathbf{x}_{it}, y_{it}) : t = 1, 2, ...\}$ is strong mixing with mixing coefficients $\alpha_i(\cdot)$. $\alpha(\cdot) \equiv \max_i \alpha_i(\cdot)$ satisfies $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $\rho \in (0, 1)$. $\{\mathbf{x}_i, \mathbf{y}_i\}$ are independent across i. $\mathbb{E}(\varepsilon_{it}) = 0$ and $\mathbb{E}(\mathbf{x}_{it}\varepsilon_{it}) = 0$ for each i and t.*

(ii) *There exist two constants c_1 and c_2 such that $0 < c_1 \leq \min_{1\leq k\leq K} \mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0} \mathbb{E}(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i)\right)$ and $\max_{1\leq i\leq N} \mu_{\max}\left(\frac{1}{T}\mathbb{E}(\mathbf{x}_i'\mathbf{x}_i)\right) \leq c_2 < \infty$.*

(iii) *There exists a constant $c_3 < \infty$ such that $\max_{i,t} \mathbb{E}\|\mathbf{x}_{it}\|^{2q} < c_3$ and $\max_{i,t} \mathbb{E}|\varepsilon_{it}|^{2q} < c_3$ for some $q > 4$.*

(iv) *$T \to \infty$. For $k = 1, ..., K$, N_k either passes to infinity or stays fixed as $T \to \infty$, and $N = O(T^2)$.*

Assumption A2. *$p_\lambda(\cdot)$ is symmetric function and is nondecreasing and concave on $[0,\infty)$. $\rho'_\lambda(s)$ exists and is continuous except for a finite number of s and $\rho'_\lambda(0+)=1$. There exists a constant $a>0$ such that $\rho_j(s)$ is constant for all $|s|\geq a\lambda$.*

Assumption A3. (i) $K=o(T/(\ln T)^2)$ and $b_{NT}\gg \ln T\sqrt{K/T}$.

(ii) *The tuning parameters λ_1 and λ_2 satisfy the following conditions: $b_{NT}\gg a\max\{\lambda_1,\lambda_2\}$, $1\gg\lambda_1\gg\frac{\ln T}{N\sqrt{T}}$, and $1\gg\lambda_2\gg\frac{\ln T}{NN_{\min}\sqrt{T}}\sqrt{\max_{1\leq k\leq K}\phi_k}$, where $N_{\min}=\min\{N_1,...,N_K\}$.*

Assumption A4. (i) *For each $k=1,...,K$, $\bar{\Phi}_k\equiv\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^T\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}'_{it}\xrightarrow{P}\Phi_k>0$ as $(N_k,T)\to\infty$ or $T\to\infty$ alone.*

(ii) *For each $k=1,...,K$, $\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^T\tilde{\mathbf{x}}_{it}\varepsilon_{it}-\mathbb{B}_{kNT}\xrightarrow{D}N(0,\Psi_k)$ as $(N_k,T)\to\infty$ or $T\to\infty$ alone where $\mathbb{B}_{kNT}=\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^T\mathbb{E}(\tilde{\mathbf{x}}_{it}\varepsilon_{it})$ is either 0 or $O(\sqrt{N_k/T})$ depending on whether $\mathbf{x}_{it}$ is strictly exogenous.*

Assumption A1(i) imposes conditions on $\{(\mathbf{x}_{it},y_{it})\}$. We require $\{(\mathbf{x}_{it},y_{it})\}$ to be weakly dependent (strong mixing is assumed here) but not necessarily stationary in the time dimension, and independent but not necessarily identically distributed in the cross section dimension. The regressor $\mathbf{x}_{it}$ can be either strictly exogenous or sequentially exogenous. Note that A1(i) does not rule out serial correlation among $\{\varepsilon_{it},t=1,2,...\}$ or $\{\mathbf{x}_{it}\varepsilon_{it},t=1,2,...\}$. A1(ii) requires that the minimum eigenvalue of $\frac{1}{TN_k}\sum_{i\in G_k^0}\mathbb{E}(\tilde{\mathbf{x}}'_i\tilde{\mathbf{x}}_i)$ be bounded away from zero and the maximum eigenvalue of $\frac{1}{T}\mathbb{E}(\mathbf{x}'_i\mathbf{x}_i)$ be bounded away from infinity, uniformly in k and i, respectively. A1(iii) imposes some moment conditions on $\mathbf{x}_{it}$ and ε_{it}. In comparison with conditions 1 and 3 in KFW which require nonrandom regressors and sub-Gaussian error terms, the conditions in A1(i)-(iii) are quite weak. A1(iv) states conditions on T, N, and N_k where we allow N_k to be fixed for some groups and to pass to infinity for other groups, thereby providing some practical flexibility in group size. In contrast, SSP require that N_k passes to infinity at the same rate as N for each k.

Assumption A2 is identical to condition 2 in KFW. Following KFW, we specify $p_\lambda(\cdot)$ as the SCAD penalty function in our simulations and applications below. Assumption A3 imposes conditions on K, b_{NT}, λ_1 and λ_2. A3(i) allows the number of groups to diverge with T and the minimum difference between two group-specific coefficients to shrink to zero at a slow rate. A3(ii) specifies the ranges of speed at which λ_1 and λ_2 shrink to zero. Assumption A4 borrows from SSP and is used in studying the asymptotic distributional properties of the Panel-CARDS estimators. If $\mathbf{x}_{it}$ contains lagged dependent variables (e.g., $y_{i,t-1}$), it is well known that the fixed effects within-group (WG) estimator has asymptotic bias of order $O(1/T)$ in homogeneous dynamic panel data models. This implies that $\mathbb{B}_{kNT}=O(\sqrt{N_k/T})$ in dynamic panel data models and bias correction is required for statistical inference unless T passes to infinity faster than N_k. See SSP for detailed discussions concerning A4.

3.2 Analysis of the basic Panel-CARDS

Next we define the oracle estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. When the grouping structure in $\mathcal{G} = \{G_1^0, ..., G_K^0\}$ is known, we can utilize the information that all coefficients $\boldsymbol{\beta}_i$ within the same true group are identical to estimate $\boldsymbol{\beta}$ by minimizing $L_{NT}(\boldsymbol{\beta})$ in (2.10). The resulting estimator of $\boldsymbol{\beta}$ is denoted $\hat{\boldsymbol{\beta}}^{oracle}$. Similarly, by using the true grouping structure, we obtain the oracle estimator $\hat{\boldsymbol{\alpha}}^{oracle}$ of $\boldsymbol{\alpha}$ with a typical block given by

$$\hat{\boldsymbol{\alpha}}_k^{oracle} = \left(\sum_{i \in G_k^0} \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i \right)^{-1} \sum_{i \in G_k^0} \tilde{\mathbf{x}}_i' \tilde{\mathbf{y}}_i \text{ for } k = 1, ..., K. \tag{3.1}$$

The following theorem reports the asymptotic properties of the basic Panel-CARDS estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Theorem 3.1 *Suppose that Assumptions A1-A3 hold. Suppose that the preliminary estimate $\tilde{\boldsymbol{\beta}}$ and tuning parameter δ together generate a segmentation $\mathcal{B}$ admissible with the true grouping pattern with probability at least $1 - \epsilon_0$. Then with probability at least $1 - \epsilon_0 - o(K/T)$, the Panel-CARDS objective function (2.15) has a strictly local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', ..., \hat{\boldsymbol{\beta}}_N')'$ such that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/T})$.*

Theorem 3.1 parallels Theorem 6 in KFW. It shows that the basic Panel-CARDS procedure includes the oracle estimator $\hat{\boldsymbol{\beta}}^{oracle}$ as a strict local minimizer with high probability. When the preliminary estimators $\tilde{\boldsymbol{\beta}}_i$ are all consistent as in our panel setup with large T, the segmentation $\mathcal{B}$ is assured to be admissible w.p.a.1 as $T \to \infty$.[6] In this case, $\epsilon_0 \equiv \epsilon_{0T} \to 0$ and we have

$$P\left(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}\right) \to 1 \text{ as } T \to \infty.$$

Given the Panel-CARDS estimate $\hat{\boldsymbol{\beta}}$, we can obtain the estimated groups by classifying individuals with the same coefficient estimate $(\hat{\boldsymbol{\beta}}_i)$ into the same group. We use $\hat{G}_k$, $k = 1, 2, ..., \hat{K}$ to denote the $\hat{K}$ estimated groups, and $\hat{\boldsymbol{\alpha}}_k, k = 1, 2, ..., \hat{K}$, to denote the group-specific estimated slope coefficients. By definition,

$$\hat{G}_k = \left\{ i \in \{1, 2, ..., N\} : \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_k \right\} \text{ for } k = 1, 2, ..., \hat{K}. \tag{3.2}$$

The following theorem reports the asymptotic distributional properties of $\hat{\boldsymbol{\alpha}}_k$.

Theorem 3.2 *Suppose that the conditions in Theorem 3.1 are satisfied. Suppose that Assumption A4 holds and $\epsilon_0 \equiv \epsilon_{0T} \to 0$ as $T \to \infty$. Then, after suitable relabeling of the indices of the true groups, we have:*

(i) $P\left(\hat{K} = K\right) \to 1$ and $P\left(\hat{G}_1 = G_1^0, ..., \hat{G}_K = G_K^0\right) \to 1$ as $T \to \infty$;

(ii) for $k = 1, \cdots, K$, $\sqrt{N_k T}(\hat{\boldsymbol{\alpha}}_k - \boldsymbol{\alpha}_k^0) - \bar{\Phi}_k^{-1} \mathbb{B}_{kNT} \xrightarrow{D} N(0, \Phi_k^{-1} \Psi_k \Phi_k^{-1})$ as either $(N_k, T) \to \infty$ or $T \to \infty$.

[6] See Theorem 3 in KFW for a proof.

Theorem 3.2(i) indicates that w.p.a.1 we can determine the correct number of groups. Theorem 3.2(ii) reports the asymptotic distribution of the group-specific estimator. As SSP remark, the oracle estimator $\hat{\boldsymbol{\alpha}}_k^{oracle}$ satisfies

$$\sqrt{N_k T}\left(\hat{\boldsymbol{\alpha}}_k^{oracle} - \boldsymbol{\alpha}_k^0\right) - \bar{\Phi}_k^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N\left(0, \Phi_k^{-1}\Psi_k\Phi_k^{-1}\right) \text{ as } (N_k, T) \to \infty \text{ or } T \to \infty$$

under Assumption A4. Theorem 3.2(ii) indicates that the Panel-CARDS estimator $\hat{\boldsymbol{\alpha}}_k$ achieves the same limit distribution as this oracle estimator with knowledge of the exact membership of each individual. In this sense, we say that Panel-CARDS estimators $\{\hat{\boldsymbol{\alpha}}_k\}$ have the asymptotic oracle property.

Given the estimated grouping structure $\left\{\hat{G}_k\right\}$, we can define the post Panel-CARDS estimator of $\boldsymbol{\alpha}_k$ as

$$\hat{\boldsymbol{\alpha}}_{\hat{G}_k} = \left(\sum_{i\in\hat{G}_k} \tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1} \sum_{i\in\hat{G}_k} \tilde{\mathbf{x}}_i'\tilde{\mathbf{y}}_i, \ k = 1, ..., \hat{K}. \tag{3.3}$$

The following theorem reports the asymptotic distribution of $\hat{\boldsymbol{\alpha}}_{\hat{G}_k}$.

Theorem 3.3 *Suppose that the conditions in Theorem 3.2 are satisfied. Then, for* $k = 1, \cdots, K$, $\sqrt{N_k T}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0) - \bar{\Phi}_k^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N(0,\ \Phi_k^{-1}\Psi_k\Phi_k^{-1})$ *as* $(N_k, T) \to \infty$ *or* $T \to \infty$.

So post Panel-CARDS estimators also share the asymptotic oracle property of the Panel-CARDS estimators. It is well known that the post-Lasso estimators have less bias than the Lasso estimators and better finite sample performance than the latter. In the simulations below, we accordingly focus on the finite sample performance of the post Panel-CARDS estimates.

It is worth mentioning that in comparison with SSP who require both N_k and T to pass to infinity, the asymptotic theory here does not require $N_k \to \infty$ or $N = \sum_{k=1}^K N_k \to \infty$. In the special case where N_k is fixed, $\mathbb{B}_{kNT} = O(\sqrt{1/T}) = o(1)$ and no bias correction is needed for either the Panel-CARDS estimators or their post-Lasso version.

3.3 Analysis of the advanced Panel-CARDS

The advanced Panel-CARDS method is an extension of basic Panel-CARDS. With some minor abuse of notation, we continue to use $\hat{\boldsymbol{\beta}}$ to denote the advanced Panel-CARDS estimator. The following theorem reports the asymptotic properties of $\hat{\boldsymbol{\beta}}$.

Theorem 3.4 *Suppose that Assumptions A1-A3 hold. Suppose that the preliminary estimate* $\tilde{\boldsymbol{\beta}}$, *the tuning parameter* δ, *and the choice of* R *together generate an admissible segmentation net* $\mathcal{N}$ *with probability at least* $1 - \epsilon_1$. *Then with probability at least* $1 - \epsilon_1 - o(K/T)$, *the Panel-CARDS objective function (2.18) has a strictly local minimizer* $\hat{\boldsymbol{\beta}}$ *such that* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}$ *and* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/T})$.

The above theorem shows that the advanced Panel-CARDS procedure includes the oracle estimator $\hat{\boldsymbol{\beta}}^{oracle}$ as a strict local minimizer with high probability. When the preliminary estimators $\tilde{\boldsymbol{\beta}}_i$ are all consistent as in our panel setup with large T, the segmentation $\mathcal{B}$ can be assured to be admissible w.p.a.1 as $T \to \infty$. In this case, $\epsilon_1 \equiv \epsilon_{1T} \to 0$ and we have $P\left(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}\right) \to 1$ as $T \to \infty$. Then analogous results as in Theorems 3.2-3.3 hold for the advanced Panel-CARDS estimators and their post-Lasso version. For brevity, we do not state the corresponding theorems.

4 Monte Carlo Simulations

In this section we conduct a small set of Monte Carlo simulations to demonstrate the finite sample performance of Panel-CARDS. We choose experimental design settings for the Monte Carlo study that enable comparisons between the basic and advanced Panel-CARDS procedures and that reflect the type of challenges likely to be present in applied work.

4.1 Data generating processes

We consider four data generating processes (DGPs).

DGP 1. Both the fixed effects μ_i and the error terms follow the i.i.d. standard normal distribution across time and individuals and are mutually independent of each other. Individuals are divided into three groups with $N_1 : N_2 : N_3 = 4 : 3 : 3$. The observations $(y_{it}, \mathbf{x}_{it})$ are generated from the panel model (2.1) where $\mathbf{x}_{it} = (x_{it1}, x_{it2})'$, $x_{it1} = 0.2\mu_i + e_{it1}$, $x_{it2} = 0.2\mu_i + e_{it2}$, e_{it1} and e_{it2} are both i.i.d. standard normal. The true coefficients are

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix} \right).$$

Note that for the first regressor, its slope coefficient is homogeneous across groups 1 and 2; and similarly for the second regressor, its slope coefficient is homogeneous across groups 2 and 3. In this case, we cannot construct an admissible segmentation using the rank of the estimates of one single slope coefficient. We want to evaluate the performance of basic Panel-CARDS and make comparisons with advanced Panel-CARDS.

DGP 2. Here we use DGP 1 in SSP. Individuals are also divided into three groups with $N_1 : N_2 : N_3 = 4 : 3 : 3$. The observations $(y_{it}, \mathbf{x}_{it})$ are generated from the panel model (2.1) where $\mathbf{x}_{it} = (x_{it1}, x_{it2})'$, $x_{it1} = 0.2\mu_i + e_{it1}$, $x_{it2} = 0.2\mu_i + e_{it2}$, e_{it1} and e_{it2} are both i.i.d. standard normal. The true coefficients are

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left(\begin{bmatrix} 0.4 \\ 1.6 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1.6 \\ 0.4 \end{bmatrix} \right).$$

DGP 3. In this DGP, we set the true number of groups to 8 where the first group has 30% of individuals and each of the other seven groups has 10% of individuals. We let $p = 2$, and the

regressors are generated as DGP 1. The true group-specific parameters take the values

$$\left(\begin{bmatrix} -4 \\ 4 \end{bmatrix}, \begin{bmatrix} -3 \\ 3 \end{bmatrix}, \begin{bmatrix} -2 \\ 2 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \end{bmatrix}, \begin{bmatrix} 3 \\ -3 \end{bmatrix}, \begin{bmatrix} 4 \\ -4 \end{bmatrix} \right).$$

DGP 4. Here we consider a dynamic panel data model where there are 3 groups with $N_1 : N_2 : N_3 = 4 : 3 : 3$. The regressors are $\mathbf{x}_{it} = (y_{i,t-1}, x_{it1}, x_{it2})'$, where (x_{it1}, x_{it2}) are generated as DGP 1. In generating T periods of observations for individual i, we first generate $T + 100$ observations with initialization $y_{i0} = 0$, and then take the last T periods of observations. The true parameter values are

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left(\begin{bmatrix} 0.6 \\ 1.5 \\ -1 \end{bmatrix}, \begin{bmatrix} 0.6 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.6 \\ 0.5 \\ 1 \end{bmatrix} \right).$$

In DGP 2-4, the fixed effects and the error terms in (2.1) are generated as in DGP 1. We will consider $N = 100$, 200 and $T = 10$, 20, 40 and 80. Since Panel-CARDS is computationally intensive, we fix the number of replications to 100 for all scenarios in this investigation.

4.2 Implementation and evaluation

For DGP 1 we use both the basic and advanced Panel-CARDS methods together with the hierarchical clustering setup. Since the performance of the basic Panel-CARDS is not robust and leads to rather unsatisfactory performance in DGP 1, we only implement advanced Panel-CARDS in DGPs 2-4. Recall that η controls the minimum percentage of observations within each estimated group. We set $\eta = 10\%, 5\%, 2\%$, and 0 to estimate the model and obtain the grouping results. When $\eta = 0$, we allow the minimum number of elements in an estimated group to be 1. The larger the value of η, the larger the number of elements for the smallest estimated group that is allowed and the smaller the number of groups estimated. For DGPs 1-2, we consider all candidate values of $\eta : 10\%, 5\%, 2\%$, and 0; for DGPs 3-4, we consider $\eta = 5\%, 2\%$, and 0 because $\eta = 10\%$ is a strong assumption when we have 8 groups in DGP 3.

The hierarchical clustering is carried out as follows.

- Let $N^* = N\eta$. For a Panel-CARDS classification $\mathcal{A}^0 = \{A_1, A_2, ..., A_{\hat{K}^0}\}$, if $|A_k| > N^*$, we consider A_k as a properly identified group; otherwise, we treat it as misclassified. Without loss of generality, we assume the properly identified groups are given by $\mathcal{A} = \{A_1, A_2, ..., A_{\hat{K}}\}$, and the misclassified members are in set $\mathcal{J} = \cup_{s=\hat{K}+1}^{\hat{K}^0} A_s$. For all members in the misclassified groups, we re-run the classification.

- For each $j \in \mathcal{J}$, we estimate its group membership by

$$k^* = \arg \min_{k \in \{1,2,...,\hat{K}^0\}; \boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_{\hat{K}}} \frac{1}{2NT} \sum_{l=1}^{\hat{K}} \sum_{i \in A_l} \sum_{t=1}^{T} \left[(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}' \boldsymbol{\beta}_l)^2 + (\tilde{y}_{jt} - \tilde{\mathbf{x}}_{jt}' \boldsymbol{\beta}_k)^2 \cdot 1\{k = l\} \right].$$

Now we re-classify the element j to group A_{k^*} for $k^* \in \{1, ..., \hat{K}\}$. In other words, we treat j as a new observation, and reclassify it to the group which makes the objective function the smallest.

- We repeat the last step for the remaining elements in $\mathcal{J}$. The final estimated grouping structure is denoted by $\widehat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, ..., \hat{G}_{\hat{K}}\}$.

We use a BIC-type information criteria to choose the tuning parameters. Given the Panel-CARDS classification results $\widehat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, ..., \hat{G}_{\hat{K}}\}$, which are obtained by using the tuning parameter vector $\boldsymbol{\lambda}$, we calculate

$$\text{IC}(\boldsymbol{\lambda}) = \ln\left(\sigma_{NT}^2(\boldsymbol{\lambda})\right) + p\hat{K}\frac{1}{2\sqrt{NT}},$$

where $\sigma_{NT}^2(\boldsymbol{\lambda}) = \frac{1}{NT}\sum_{s=1}^{\hat{K}}\sum_{i\in A_s}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\hat{\boldsymbol{\beta}}_s(\boldsymbol{\lambda}))^2$, the $\hat{\boldsymbol{\beta}}_s(\boldsymbol{\lambda})$'s are post Panel-CARDS and hierarchical clustering estimators, and here we make their dependence on $\boldsymbol{\lambda}$ explicit.

We report the frequency of obtaining a particular number of groups based on 100 replications for all DGPs. Despite the importance of correct determination of the number of groups, it does not show how similar the estimated groups are to the true groups. Following KFW, we use the normalized mutual information (NMI) measure to assess the similarity between the estimated grouping structure $\widehat{\mathcal{G}}$ and the true grouping structure $\mathcal{G}$. For two classifications/grouping structures $\mathcal{A} = \{A_1, A_2, \cdots\}$ and $\mathcal{B} = \{B_1, B_2, \cdots\}$ on the same set $\{1, 2, \cdots, N\}$, the NMI is defined as

$$\text{NMI}(\mathcal{A}, \mathcal{B}) = \frac{I(\mathcal{A}, \mathcal{B})}{\sqrt{H(\mathcal{A})H(\mathcal{B})}},$$

where

$$I(\mathcal{A}, \mathcal{B}) = \sum_{i,j}(|A_i \cap B_j|/N)\ln\left(\frac{|A_i \cap B_j|/N}{|A_i|/N \cdot |B_j|/N}\right) \quad \text{and} \quad H(\mathcal{A}) = -\sum_i \frac{|A_i|}{N}\ln\left(\frac{|A_i|}{N}\right).$$

When $\mathcal{A}$ and $\mathcal{B}$ have the same classification, we have $I(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) = H(\mathcal{B})$, and $\text{NMI}(\mathcal{A}, \mathcal{B}) = 1$. In general, the more similar two classifications are, the closer their NMI value is to 1.

We report $\text{NMI}(\widehat{\mathcal{G}}, \mathcal{G})$ for all DGPs. In addition, we report the root mean square error (RMSE) for DGP 2 only to save space.

4.3 Simulation results

Table 1 reports the frequency of the estimated number of groups for DGP 1 based on the basic Panel-CARDS (b-Panel-CARDS). Apparently, the performance of b-Panel-CARDS in DGP 1 is poor, which is as expected. Theorem 3.1 requires an admissible segmentation for the b-Panel-CARDS to work well. But the choice of the group-specific parameter values in DGP 1 rules out the possibility of admissible segmentation by using the preliminary estimates of a single coefficient to construct the segmentation.

Because the b-Panel-CARDS is not robust against certain patterns of group-specific parameter values such as those in DGP 1, below we will focus on the performance of the advanced Panel-CARDS (a-Panel-CARDS).[7] We use $R = 2$ regressors to construct the segmentation net. Given the matrix of preliminary estimates, $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \cdots, \tilde{\boldsymbol{\beta}}_N)$, we calculate the sample variance of each row of $\tilde{\mathbf{B}}$ and choose the two regressors with the largest variances for their coefficient estimates to construct the segmentations.

Table 2 and Figure 1 report the classification results for DGP 1 based on the a-Panel-CARDS for different combinations of N, T, and η. Unsurprisingly, the results in Table 2 are much better than those in Table 1. Table 2 suggests that when we set the tuning parameter η to be 10%, the a-Panel-CARDS procedure performs well even when T is very small relative to N, and we can correctly determine the number of groups with a large probability. When T increases, we have more accurate preliminary estimates of the parameters and the classification also improves. When η decreases, the a-Panel-CARDS tends to estimate more groups than the correct number of groups for small values of T; but its performance quickly improves as T increases. Figure 1 shows the NMI between the estimated group structure $\widehat{\mathcal{G}}$ and the true group structure $\mathcal{G}$ for different combinations of N, T, and η. It suggests that as T increases, the NMI between $\widehat{\mathcal{G}}$ and $\mathcal{G}$ increases rapidly. When $T = 80$, the estimation is almost as good as the oracle for all values of η. We also note that the performance of a-Panel-CARDS with $\eta = 2\%$ or 5% significantly improves that with $\eta = 0$, but a further increase of η does not necessarily lead to improved performance.

Table 3 reports the frequency of the estimated number of groups for DGP 2 based on the a-Panel-CARDS. It suggests that when T is small (10 or 20), a higher value of η helps considerably in determining the correct number the groups as in DGP 1. But when T is sufficiently large (say, 80), the a-Panel-CARDS with $\eta = 0$ can also achieve almost perfect classification. Comparing Table 3 with the results of DGP 1 in SSP, we find that the performance here is not as good as theirs. But note here that we use the a-Panel-CARDS, whose number of penalty terms approximately doubles that of the b-Panel-CARDS approach. As remarked earlier, increasing penalty terms has the side effect of accumulating errors. When we use the b-Panel-CARDS (which is sufficient for DGP 2) and set $\eta = 10\%$, its performance is comparable to that of SSP and significantly dominates the latter when $T = 10$.

Figure 2 reports the NMI for DGP 2 for various combinations of N, T, and η. The NMI patterns in Figure 2 are similar to those in Figure 1 for DGP 1. Figure 3 presents the RMSE of for different combinations of N, T, and η. Also reported in the figure is the RMSE for the estimates $\{\tilde{\beta}_i\}$ which are obtained by treating every unit (i) as a group and labeled as "unitwise" estimates. To evaluate the finite sample gains from using the a-Panel-CARDS, we compare its RMSE with that of unitwise estimators and oracle estimators. Figure 3 suggests that the a-Panel-CARDS estimators

[7] But this does not mean that the a-Panel-CARDS dominates the b-Panel-CARDS in all cases. In DGP 2 below, we find that b-Panel-CARDS can generate more accurate grouping results than the a-Panel-CARDS or SSP's C-Lasso. In real data applications, we apply both the a-Panel-CARDS and b-Panel-CARDS, and then rely on the information criteria introduced in the last subsection to choose between them. And we call the result Panel-CARDS.

Table 1: Frequency of obtaining the estimated number of groups in DGP 1 based on b-Panel-CARDS

η	N	T	1	2	3	4	5	6	7	8+
0.10	100	10	0.00	0.20	0.63	0.16	0.01	0.00	0.00	0.00
	100	20	0.00	0.03	0.87	0.10	0.00	0.00	0.00	0.00
	100	40	0.00	0.01	0.89	0.10	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	0.89	0.11	0.00	0.00	0.00	0.00
	200	10	0.00	0.25	0.61	0.13	0.01	0.00	0.00	0.00
	200	20	0.00	0.09	0.79	0.12	0.00	0.00	0.00	0.00
	200	40	0.00	0.03	0.85	0.10	0.02	0.00	0.00	0.00
	200	80	0.00	0.01	0.92	0.06	0.01	0.00	0.00	0.00
0.05	100	10	0.00	0.00	0.36	0.42	0.22	0.00	0.00	0.00
	100	20	0.00	0.00	0.55	0.37	0.07	0.01	0.00	0.00
	100	40	0.00	0.00	0.61	0.32	0.07	0.00	0.00	0.00
	100	80	0.00	0.00	0.57	0.36	0.06	0.01	0.00	0.00
	200	10	0.00	0.00	0.13	0.45	0.31	0.11	0.00	0.00
	200	20	0.00	0.00	0.23	0.41	0.26	0.07	0.03	0.00
	200	40	0.00	0.00	0.60	0.29	0.08	0.03	0.00	0.00
	200	80	0.00	0.00	0.43	0.45	0.11	0.01	0.00	0.00
0.02	100	10	0.00	0.00	0.09	0.23	0.35	0.28	0.02	0.03
	100	20	0.00	0.00	0.27	0.49	0.19	0.04	0.01	0.00
	100	40	0.00	0.00	0.56	0.38	0.06	0.00	0.00	0.00
	100	80	0.00	0.00	0.50	0.33	0.16	0.01	0.00	0.00
	200	10	0.00	0.00	0.00	0.06	0.14	0.10	0.22	0.48
	200	20	0.00	0.00	0.06	0.15	0.22	0.18	0.11	0.28
	200	40	0.00	0.00	0.12	0.35	0.29	0.17	0.04	0.03
	200	80	0.00	0.00	0.11	0.37	0.34	0.15	0.03	0.00
0	100	10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99
	100	20	0.00	0.00	0.00	0.01	0.03	0.04	0.07	0.85
	100	40	0.00	0.00	0.00	0.03	0.05	0.18	0.18	0.56
	100	80	0.00	0.00	0.00	0.02	0.11	0.19	0.27	0.41
	200	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	200	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	200	40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	200	80	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

generally outperform the unitwise estimators, and when T increases to 80, their performance is almost as good as the oracle. With respect to η, we again find that a choice of $\eta = 2\%$ or 5% tends to outperform $\eta = 0$.

Table 4 and Figure 4 show that the classification results for DGP 3 where the true number of groups is reasonably large (8 here). They show that the classification is very accurate even in this challenging scenario as long as $T \geq 20$ and $\eta \geq 2\%$. As before, the choice of $\eta = 0$ produces good classification results only when T is sufficiently large.

Table 5 and Figure 5 report the classification results for DGP 4 where the panel is a dynamic panel. Apparently, the a-Panel-CARDS performs very well in this situation unless T is very small and $\eta = 0$. The general conclusions from DGP 1-3 also hold here.

Table 2: Frequency of obtaining the estimated number of groups in DGP 1 based on a-Panel-CARDS

η	N	T	1	2	3	4	5	6	7	8+
0.10	100	10	0.00	0.04	0.84	0.11	0.01	0.00	0.00	0.00
	100	20	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.06	0.82	0.11	0.01	0.00	0.00	0.00
	200	20	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
	200	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.05	100	10	0.00	0.00	0.51	0.37	0.10	0.02	0.00	0.00
	100	20	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.49	0.37	0.11	0.03	0.00	0.00
	200	20	0.00	0.01	0.94	0.02	0.03	0.00	0.00	0.00
	200	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.02	100	10	0.00	0.00	0.34	0.31	0.23	0.06	0.05	0.01
	100	20	0.00	0.00	0.93	0.07	0.00	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.03	0.36	0.20	0.14	0.08	0.08	0.11
	200	20	0.00	0.00	0.96	0.01	0.01	0.00	0.00	0.02
	200	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0	100	10	0.00	0.00	0.00	0.00	0.01	0.05	0.06	0.88
	100	20	0.00	0.00	0.08	0.24	0.21	0.13	0.13	0.21
	100	40	0.00	0.00	0.70	0.26	0.03	0.01	0.00	0.00
	100	80	0.00	0.00	0.97	0.03	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.01	0.01	0.02	0.01	0.05	0.90
	200	20	0.00	0.03	0.05	0.07	0.12	0.17	0.13	0.43
	200	40	0.00	0.01	0.68	0.24	0.06	0.01	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

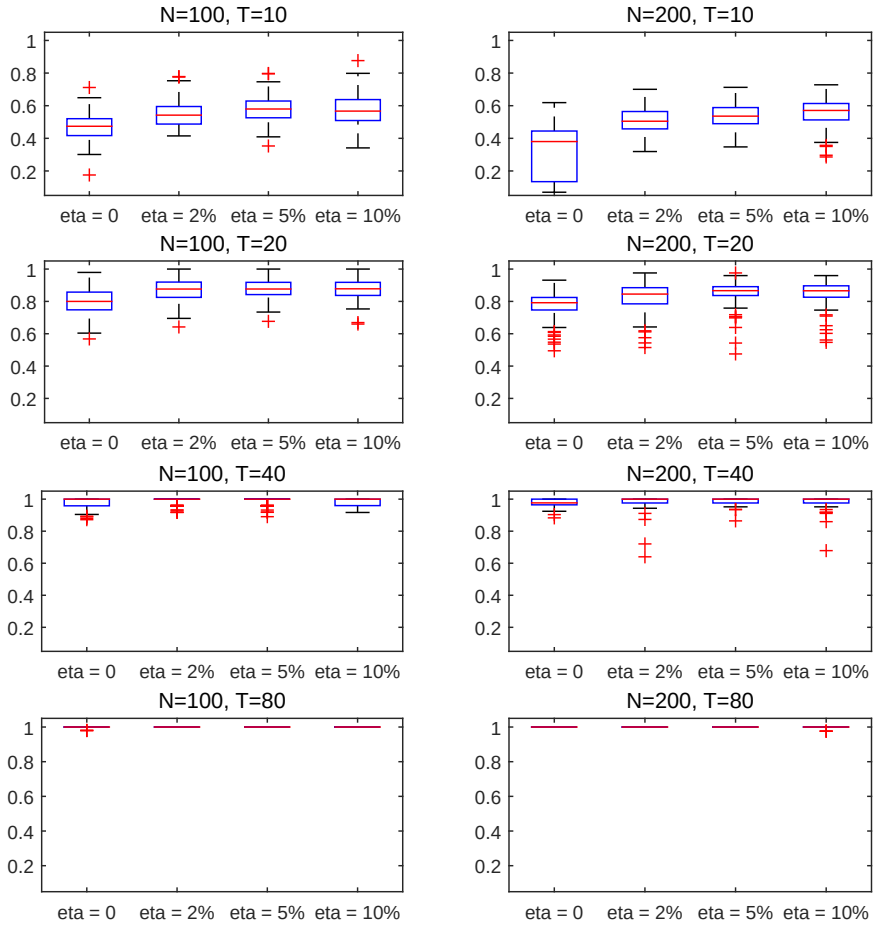

Figure 1: NMI of DGP 1 classification results using a-Panel-CARDS

5 Empirical Applications

5.1 Income and democracy

As Acemoglu et al. (2008) remark, one of the most notable empirical regularities in modern political economy is the positive relationship between income per capita and democracy. Existing studies such as Barro (1999) and Acemoglu et al. (2008) establish a strong cross-country correlation between income and democracy, but do not typically control for cross-country heterogeneity in the slope coefficients. For different countries, the relationship between the two variables might well be similar or equally well be quite different. In South Korea, the degree of democracy increases when the economy is growing steadily. Similar patterns emerge for other countries such as Japan, Spain, and Romania. However, for China the story is quite different. The democracy index composed by Freedom House has not changed very much over the last three decades or more for China despite the fact that China has made remarkable economic progress over the same period. Moreover, for some countries like Iran and Malaysia, a negative correlation is observed between income and democracy.

Table 3: Frequency of obtaining the estimated number of groups in DGP 2 based on a-Panel-CARDS

η	N	T	1	2	3	4	5	6	7	8+
0.10	100	10	0.00	0.06	0.76	0.18	0.00	0.00	0.00	0.00
	100	20	0.00	0.01	0.88	0.08	0.03	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.25	0.53	0.19	0.03	0.00	0.00	0.00
	200	20	0.00	0.17	0.51	0.29	0.01	0.02	0.00	0.00
	200	40	0.00	0.00	0.97	0.03	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.05	100	10	0.00	0.02	0.39	0.47	0.10	0.02	0.00	0.00
	100	20	0.00	0.02	0.87	0.07	0.02	0.01	0.01	0.00
	100	40	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.08	0.50	0.19	0.15	0.07	0.01	0.00
	200	20	0.00	0.16	0.41	0.14	0.11	0.07	0.05	0.06
	200	40	0.00	0.00	0.83	0.07	0.04	0.05	0.01	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.02	100	10	0.00	0.01	0.33	0.37	0.21	0.07	0.01	0.00
	100	20	0.00	0.02	0.76	0.20	0.02	0.00	0.00	0.00
	100	40	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.06	0.29	0.33	0.13	0.08	0.08	0.03
	200	20	0.00	0.16	0.42	0.14	0.04	0.08	0.10	0.06
	200	40	0.00	0.00	0.89	0.06	0.02	0.00	0.01	0.02
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0	100	10	0.00	0.00	0.00	0.00	0.00	0.02	0.06	0.92
	100	20	0.00	0.02	0.10	0.13	0.18	0.13	0.15	0.29
	100	40	0.00	0.00	0.53	0.34	0.09	0.02	0.02	0.00
	100	80	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.99
	200	20	0.00	0.00	0.05	0.05	0.08	0.04	0.13	0.65
	200	40	0.00	0.00	0.30	0.33	0.13	0.10	0.07	0.07
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

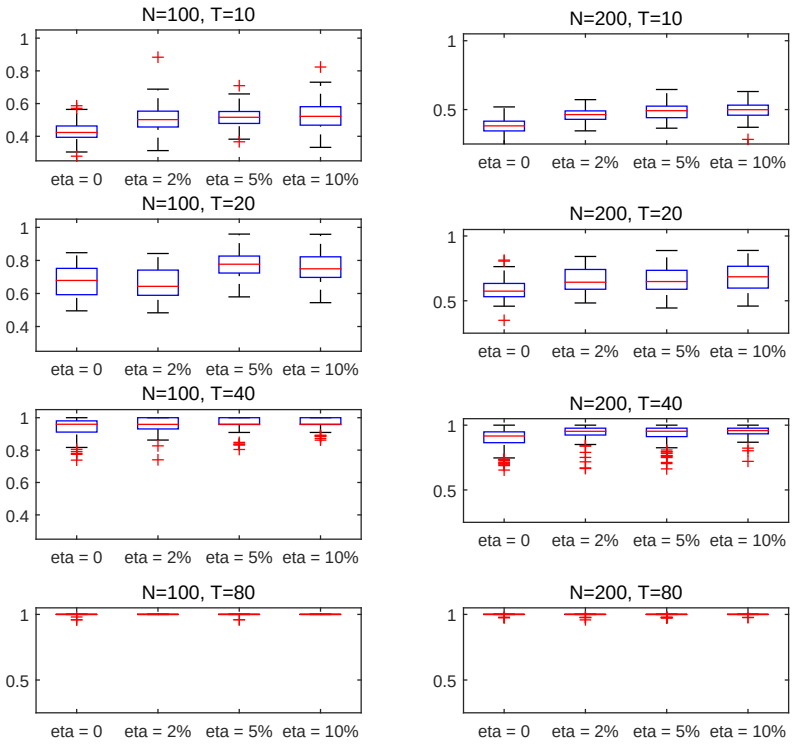

Figure 2: NMI of DGP 2 classification results using a-Panel-CARDS

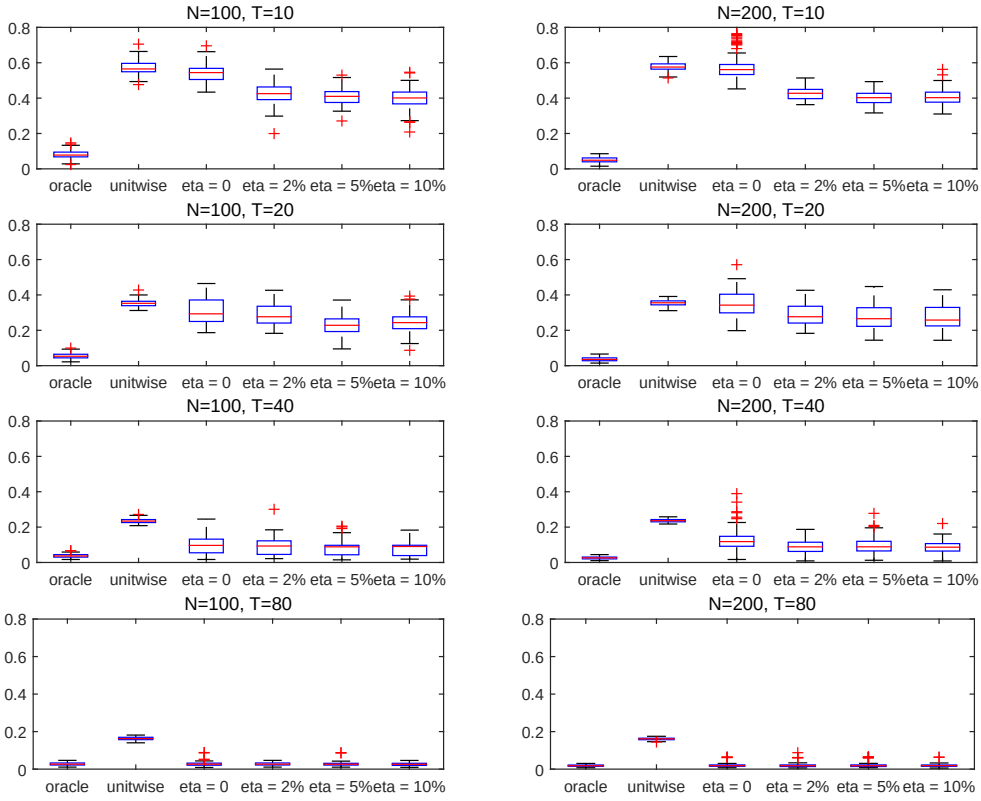

Figure 3: Root mean square error of DGP 2 post classification estimators

Table 4: Frequency of obtaining the estimated number of groups in DGP 3 based on a-Panel-CARDS

η	N	T	6	7	8	9	10	11	12	13+
0.05	100	10	0.71	0.18	0.09	0.01	0.00	0.00	0.00	0.01
	100	20	0.00	0.00	0.91	0.09	0.00	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.10	0.28	0.59	0.03	0.00	0.00	0.00	0.00
	200	20	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.02	100	10	0.00	0.00	0.34	0.31	0.23	0.06	0.05	0.01
	100	20	0.00	0.00	0.93	0.07	0.00	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.67	0.27	0.05	0.01	0.00	0.00
	200	20	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0	100	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	100	20	0.00	0.00	0.01	0.02	0.06	0.05	0.16	0.70
	100	40	0.00	0.00	0.51	0.27	0.18	0.04	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
	200	20	0.00	0.00	0.01	0.02	0.05	0.19	0.22	0.51
	200	40	0.00	0.00	0.68	0.28	0.02	0.02	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

Table 5: Frequency of obtaining the estimated number of groups in DGP 4 based on a-Panel-CARDS

η	N	T	1	2	3	4	5	6	7	8+
0.05	100	10	0.00	0.00	0.85	0.14	0.01	0.00	0.00	0.00
	100	20	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.79	0.15	0.05	0.01	0.00	0.00
	200	20	0.00	0.00	0.94	0.05	0.01	0.00	0.00	0.00
	200	40	0.00	0.00	0.98	0.02	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0.02	100	10	0.00	0.00	0.48	0.35	0.12	0.04	0.01	0.00
	100	20	0.00	0.00	0.94	0.03	0.02	0.01	0.00	0.00
	100	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	100	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.53	0.31	0.09	0.06	0.01	0.00
	200	20	0.00	0.00	0.87	0.07	0.02	0.03	0.01	0.00
	200	40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
0	100	10	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.97
	100	20	0.00	0.04	0.08	0.18	0.22	0.22	0.16	0.10
	100	40	0.00	0.01	0.91	0.07	0.01	0.00	0.00	0.00
	100	80	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00
	200	10	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.99
	200	20	0.00	0.02	0.03	0.13	0.09	0.11	0.12	0.50
	200	40	0.00	0.01	0.83	0.11	0.04	0.00	0.00	0.01
	200	80	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00

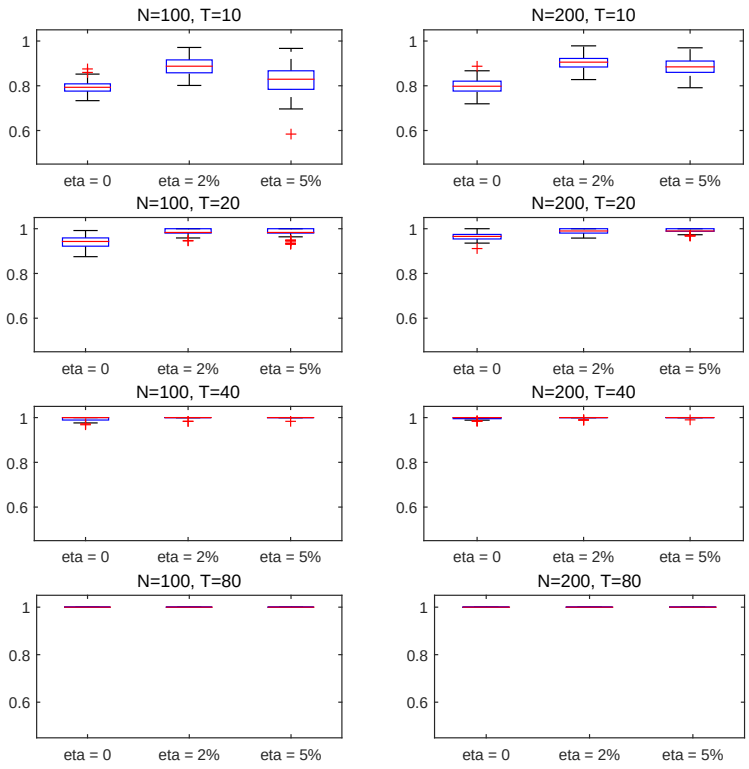

Figure 4: NMI of DGP 3 classification results using a-Panel-CARDS

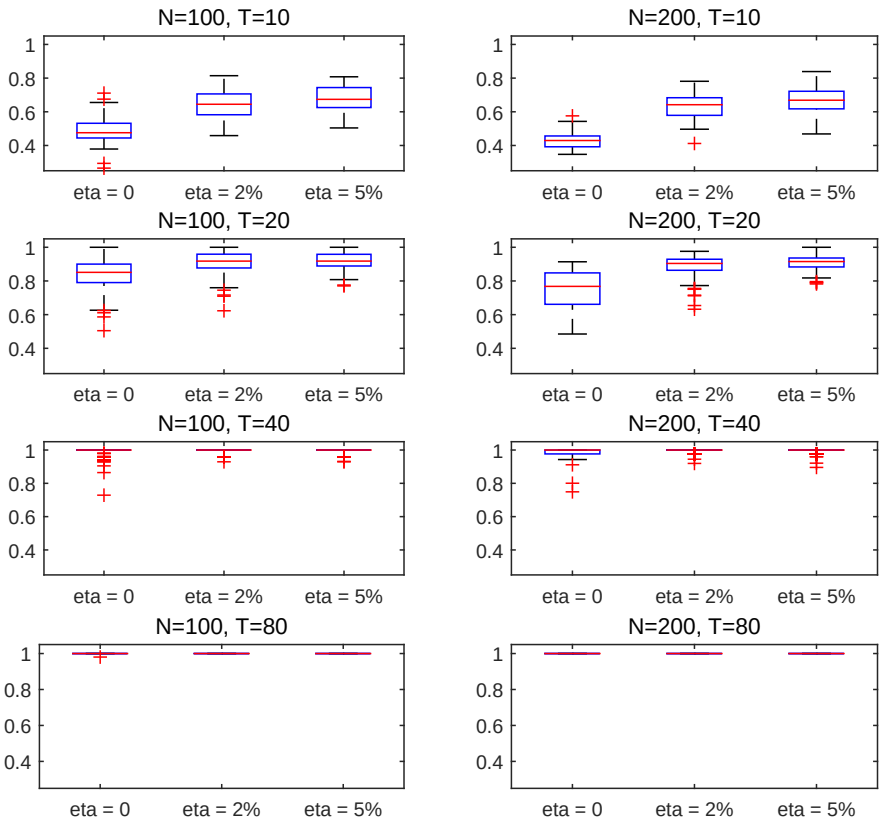

Figure 5: NMI of DGP 4 classification results using a-Panel-CARDS

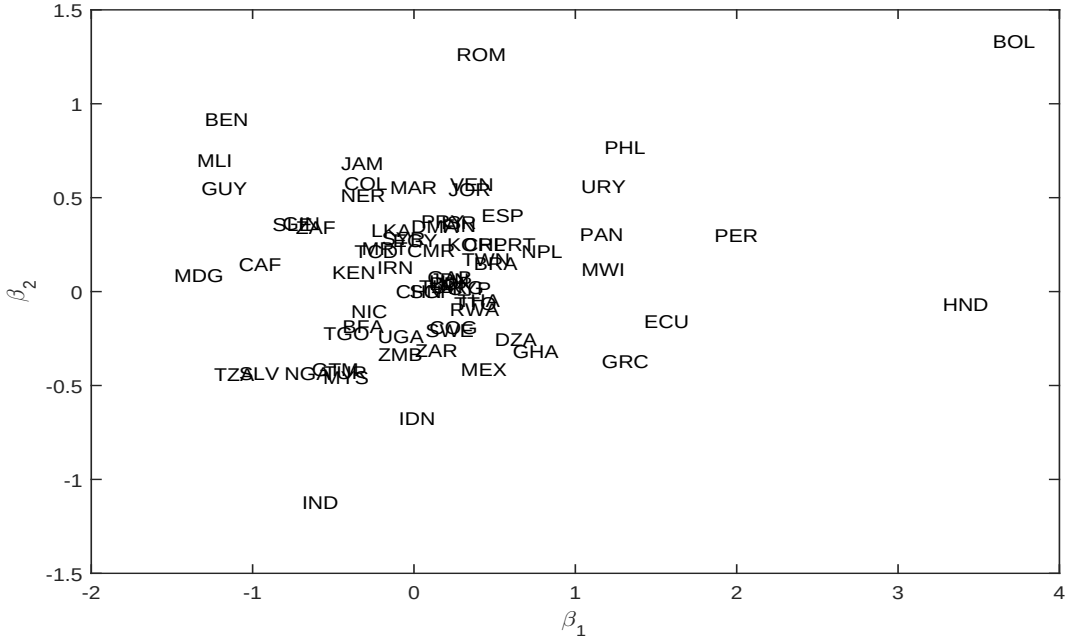

Figure 6: Scatter Plot of Preliminary Estimates

These observations motivate the use of more flexible panel modeling methods that permit some individual heterogeneity and potential country groupings of the type that are admitted within the latent panel structure model studied in this paper.

Following the lead of Acemoglu et al. (2008) and Bonhomme and Manresa (2015), we consider the following regression model

$$d_{it} = \beta_{i1} I_{i,t-1} + \beta_{i2} d_{i,t-1} + \mu_i + \varepsilon_{it}, \ i = 1, ..., N, \ t = 1, ..., T, \tag{5.1}$$

where d_{it} denotes a measure of democracy for country i in period t, I_{it} denotes the logarithm of the real GDP per capita for country i in period t, μ_i is the fixed effect, ε_{it} is the error term, and β_{i1} and β_{i2} are the slope coefficients, which are assumed to be constant across countries in early studies. See Acemoglu et al. (2008) and Bonhomme and Manresa (2015) for detailed descriptions of the variables d_{it} and I_{it}. As in these latter papers, we use a balanced panel dataset where the number of countries (N) is 74 and the time index t runs from 1 to 7. Here each time period corresponds to a five-year interval over the period 1961-2000. For example, $t = 0$ refers to the 1961-1965 period.

Without assuming any latent group structure, we can estimate the model in (5.1) by minimizing the non-penalized objective function in (2.10). Let $(\tilde{\beta}_{i1}, \tilde{\beta}_{i2})'$ denote the estimates. Since $T = 7$ is relatively small, these estimates cannot be very accurate. To get an intuitive idea about these preliminary estimates, we display their scatter plot in Figure 6. From this figure we see that these estimates have wide dispersion over the plane from which it is hard to discern any pattern.

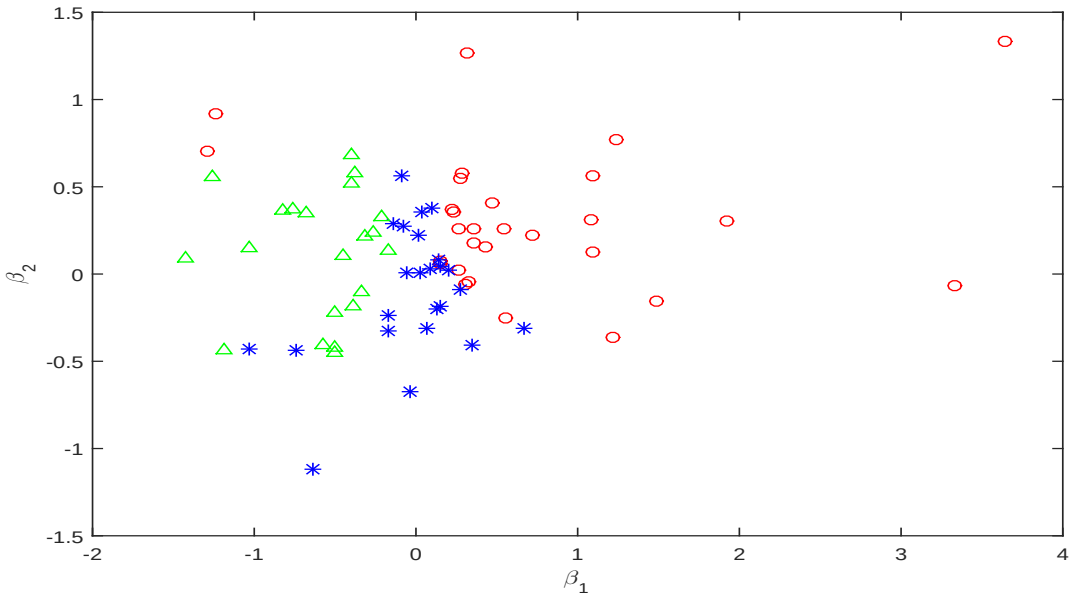

Figure 7: Scatter Plot of Classification Results

Next, we apply Panel-CARDS to determine the number of groups and estimate the group-specific parameters. We assume that each group contains at least $\eta = 2\%$ of the countries and apply the IC to choose the tuning parameter as in the simulations. The classification results are reported in Table 6 and Figure 7. Table 6 suggests that we can identify three groups and each group contains a fairly large number of countries. To connect with Table 6, we denote green triangles for group 1, blue stars for group 2, and red circles for group 3. The differences among these three groups are significant.

Table 7 reports the estimation results for each group-specific parameter and those for the pooled fixed effects (FE) estimates, all of which are bias-corrected by using the half-panel jackknife of Dhaene and Jochmans (2015). The last column in Table 7 reports the long run effect (LRE) of income on democracy: $\beta_1/(1-\beta_2)$. Based on these estimates of the effect of income on democracy, we classify countries into three groups: Group 1 is a "negative effect" group, Group 2 a "small effect" group, and Group 3 a "large effect" group. Thus, income has a negative association-effect on democracy in Group 1, a small positive association-effect on democracy in Group 2, and a large positive association-effect on democracy in Group 3.

These group selections and empirical results obtained by Panel-CARDS estimation can be compared with full panel regression outcomes. If we pool all countries together and estimate a homogeneous panel, the findings show only a positive association-effect of income with democracy, an outcome that fails to explain the disparate country phenomena discussed at the beginning of this section.

Table 6: Classification Results of Countries/Regions

Group 1: "negative effect" group ($\|\hat{G}_1\| = 21$)				
Burkina Faso	Central African Rep.	Chad	Colombia	Guatemala
Guinea	Guyana	Iran	Jamaica	Kenya
Madagascar	Malaysia	Mauritania	Nicaragua	Niger
Sierra Leone	South Africa	Sri Lanka	Tanzania	Togo
Turkey				
Group 2: "small effect" group ($\|\hat{G}_2\| = 24$)				
Argentina	Burundi	Cameroon	China	Congo Dem. Rep.
Congo Rep.	Dominican Rep.	Egypt Arab Rep.	El Salvador	Gabon
Ghana	India	Indonesia	Mexico	Morocco
Nigeria	Paraguay	Rwanda	Singapore	Sweden
Syrian Arab Rep.	Tunisia	Uganda	Zambia	
Group 3: "large effect" group ($\|\hat{G}_3\| = 29$)				
Algeria	Benin	Bolivia	Brazil	Chile
Cyprus	Ecuador	Finland	Greece	Honduras
Israel	Japan	Jordan	Korea Rep.	Luxembourg
Malawi	Mali	Nepal	Panama	Peru
Philippines	Portugal	Romania	Spain	Taiwan
Thailand	Trinidad and Tobago	Uruguay	Venezuela RB	

Table 7: Regression Results

	β_1			β_2			LRE
	estimates	s.e.	t-stat	estimates	s.e.	t-stat	
Group 1 ("negative effect")	-0.416	0.068	-6.134	0.179	0.061	2.939	-0.507
Group 2 ("small effect")	0.248	0.017	8.200	-0.013	0.079	-0.232	0.245
Group 3 ("large effect")	0.392	0.052	7.502	0.507	0.069	7.314	0.796
Pooled FE model	0.076	0.017	2.912	0.492	0.048	10.362	0.151

5.2 Minimum wage and unemployment

The relationship between minimum wage and unemployment has been widely studied in labor economics; see Brown (1999) for a summary. Conventional economic theory suggests that a rise in the minimum wage should lead to reduced employment and thus a higher unemployment rate. This assertion is challenged by empirical evidence in different ways, depending on the methodological approach employed. As Dube et al. (2010) remark, the minimum wage literature in the United States can be classified into two categories. One is based on traditional national level studies, and the other is based on case studies. National level studies such as Neumark and Washer (1992, 2007) use all cross-state variation in the minimum wage over time to estimate the employment effects of increase in minimum wage. Case studies such as Card and Krueger (1994, 2000) and Neumark and Wascher (2000) typically compare adjoining local areas with different minimum wages around the time of a policy change. In both kinds of study, the conclusions are mixed. For example, Card and Krueger (1994) study the impact of a minimum wage rise on employment using survey data for 410 fast-food restaurants in New Jersey and Eastern Pennsylvania and find that an increase in the minimum wage causes an increase in employment. In contrast, Neumark and Wascher (2000) re-examine the issue for the same two states by using administrative payroll data but find negative effects of a minimum wage rise on employment. Dube et al. (2010) show that both approaches may generate misleading results when unobserved heterogeneity is not properly accounted for.

Given these mixed findings concerning the effect of the minimum wage on employment, we might conjecture that unobserved slope heterogeneity in the across-state data is partly responsible for the mixed evidence. The panel structure model is designed to cope with unobserved heterogeneity in the response function and this motivates the use of the following modeling framework to accommodate potential heterogeneity

$$ur_{it} = \beta_{1i} ur_{i,t-1} + \beta_{2i} gr_{i,t-1} + \beta_{3i} mw_{i,t-1} + \mu_i + \varepsilon_{it}, \tag{5.2}$$

where ur_{it}, gr_{it} and mw_{it} denote the unemployment rate, GDP growth rate, and real minimum wage rate (deflated by the CPI)[8] for state i in year t, respectively, μ_i is a fixed effect, ε_{it} is an error term, and $\{\beta_{1i}, \beta_{2i}, \beta_{3i}\}$ denote heterogenous slope response parameters that may have certain latent group structures. We use US panel data for all 50 states from 1988 to 2014. So $N = 50$ and $T = 26$ in our study. All data are downloaded from the Bureau of Labor Statistics and the Federal Reserve Bank of St. Louis. We normalize the four variables to have mean 0 and variance 1 for all states.

Implementing the a-Panel-CARDS procedure, we obtain the classification results and post classification estimates reported in Tables 8 and 9, respectively. Table 8 suggests that the 50 states can be classified into two groups, each group containing roughly one half of the states. Table 9 reports the group-specific estimation results together with the pooled FE estimation results, where

[8] For most states, there are state minimum wage and federal minimum wage rates. We take the higher one as the state minimum wage.

Table 8: Classification Results of States

Group 1: "positive effect" group ($\|\hat{G}_1\| = 27$)				
Alabama	Arizona	California	Colorado	Connecticut
Florida	Georgia	Hawaii	Illinois	Maine
Maryland	Massachusetts	Michigan	Nevada	New Hampshire
New Jersey	New York	North Carolina	Ohio	Pennsylvania
Rhode Island	South Carolina	Texas	Utah	Virginia
Washington	Wisconsin			
Group 2: "negative effect" group ($\|\hat{G}_2\| = 23$)				
Alaska	Arkansas	Delaware	Idaho	Indiana
Iowa	Kansas	Kentucky	Louisiana	Minnesota
Mississippi	Missouri	Montana	Nebraska	New Mexico
North Dakota	Oklahoma	Oregon	South Dakota	Tennessee
Vermont	West Virginia	Wyoming		

all estimated are bias-corrected via the half-panel jackknife. The estimates of the coefficients of the lagged dependent variable are similar across Groups 1 and 2. For each group, the impact of GDP growth on the unemployment rate is strongly negative, which accords with Okun's law. The pooled FE estimation results suggest that increases in the minimum wage have barely any effect on the unemployment rate. The group results differ significantly: in Group 1, we find that an increasing minimum wage leads to a higher unemployment rate; but in Group 2, an increase in minimum wage causes a drop in the unemployment rate. For both groups, the coefficients are statistically significant at the 10% level, but they cancel each other out in the pooled FE estimation.

Naturally, it is interesting to contemplate reasons for these observed group differences in state outcomes. To provide some intuition, we present the geographic distribution of the classification results, and mark them on the map in Figure 8. States classified in Groups 1 and 2 are painted blue and white, respectively. Although our methodology makes no use of geographic information, the map shows that the observed geographic pattern is surprisingly regular. Almost all Group 2 (colored white) states are connected and located in the middle region of the United States. Group 1 (colored blue) states are largely scattered around the east and west coasts of the United States.

This map pattern is naturally reminiscent of the standard geopolitical map of American politics involving so-called blue states and red states.[9] In addition, by sorting the 2014 GDP outcomes from largest to smallest for the 50 states, we find that the top 15 largest economy states are all included in Group 1 (blue). One possible explanation is as follows: due to geographic, historical, transportation, demographic, and natural resource differences, people from the states in Group 1 and Group 2 have different employment choice sets, different networking opportunities, different exposure to the various manufacturing, mining, technological, educational, and financial industries, as well as different political opinions. Exploring the underlying determinants of these socio-economic-political differences is clearly of substantial interest in economic-political geography but is beyond the scope of the current paper.

[9] For example, readers may refer to https://en.wikipedia.org/wiki/United_States_presidential_election,_2012

Table 9: Regression Results

	β_1			β_2			β_3		
	estimates	s.e.	t-stat	estimates	s.e.	t-stat	estimates	s.e.	t-stat
Group 1	0.617	0.029	21.258	-0.424	0.034	-12.455	0.059	0.033	1.748
Group 2	0.864	0.040	21.516	-0.175	0.031	-5.577	-0.080	0.043	-1.855
Pooled FE	0.765	0.025	30.094	-0.295	0.025	-11.934	-0.010	0.028	-0.367

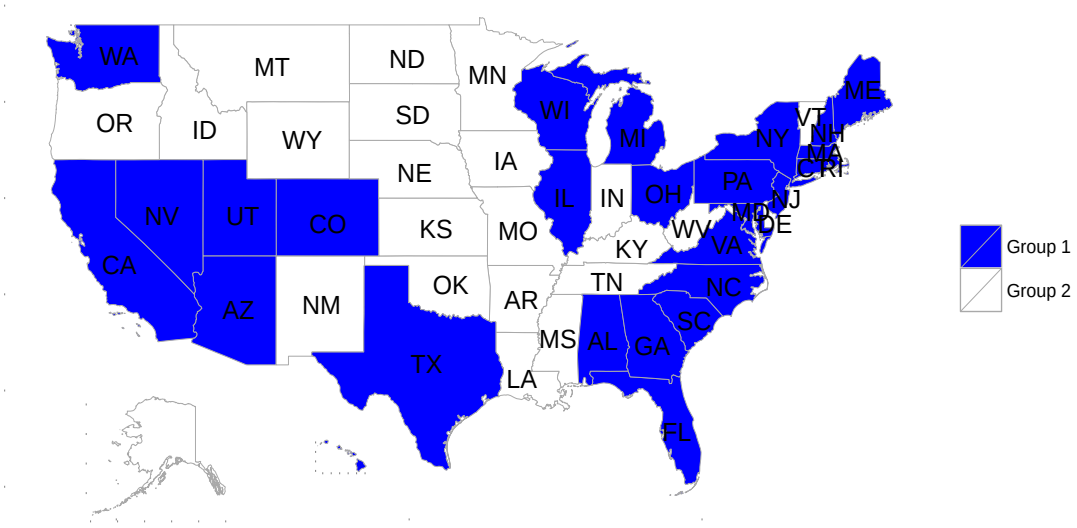

Figure 8: Classification Results on the Map

6 Conclusion

Panel data offer empirical investigators the opportunity to study individual unit behavior over time which provides the appealing prospect of increased precision in estimation due to cross section averaging. But this advantage hinges on the validity of homogeneous responses in the individual units to system covariates and the predetermined variables. Assessing the validity of such homogeneous response conditions is an important feature of successful panel data research. When homogeneity is absent and further information is lacking, empirical research is inevitably reliant on econometric methodology to assist in discovering any latent structures in the data which may lead to homogeneous sub-classes wherein cross section averaging will be valid and effective.

This paper combines with other recent work in providing such methodology for the discovery and estimation of latent structures in panel data. Our approach extends to a systematic panel framework some recent research on the CARDS method proposed by KFW. The Panel-CARDS procedure developed here is data-driven and enables identification and estimation of latent group structures compatible with oracle estimation without the use of auxiliary variates to achieve empirical classification. In comparison with the CARDS method, we consider the slope parameters of each individual unit as a whole rather than as a special case of a cross section model. Together with the use of a new concept of controlled classification of multidimensional quantities called the segmentation net, this framework provides a robust approach to group selection. If prior information about the minimum number of elements in each group does happen to be available, the method also allows for hierarchical clustering to improve estimation accuracy.

We apply the new Panel-CARDS methodology to revisit two longstanding examples of panel data research in economics. Our study of the international relationship between income and democracy identifies three latent groups of countries which demonstrate distinctive association effects, each relating income to democracy in a different way. Our study of the effect of minimum wage legislation on unemployment rates in the United States identifies two latent groups within the 50 American states, one in which the unemployment rate responds negatively to increases in the minimum wage and a second group where the response is positive. These applications demonstrate that it is possible to take advantage of increased precision in estimation from cross section averaging while at the same time identifying those subgroups of a panel in which homogeneous responses are validated by the data.

References

Acemoglu, D., Johnson, S., Robinson, J. A., and Yared, P. (2008), Income and democracy, *American Economic Review*, 98, 808-842.

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

Barro, R. J. (1999), Determinants of democracy, *Journal of Political Economy*, 107, 158-183.

Bester, C. A., Hansen, C. B. (2016), Grouped effects estimators in fixed effects models, *Journal of Econometrics*, 90, 197-208.

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

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

Brown, C. (1999), Minimum wages, employment, and the distribution of income, *Handbook of Labor Economics*, 3, 2101-2163.

Browning, M. and Carro, J. (2007), Heterogeneity and microeconometrics modeling, *Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society*, pp. 45-74.

Browning, M. and Carro, J. (2010), Heterogeneity in dynamic discrete choice models, *Econometrics Journal*, 13, 1-39.

Bühlmann, P. and van der Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Springer.

Card, D. and Krueger, A. B. (1994), Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania, *American Economic Review*, 84, 772-793.

Card, D. and Krueger, A. B. (2000), Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply, *American Economic Review*, 90, 1397-1420.

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

Dube, A., Lester, T. W., and Reich, M. (2010), Minimum wage effects across state borders: estimates using contiguous counties, *Review of Economics and Statistics*, 92, 945-964.

Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96, 1348-1360.

Fan, J., Lv, J., and Qi, L. (2011), Sparse high dimensional models in economics, *Annual Review of Economics*, 3, 291-317.

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

Kasahara, H. and Shimotsu, K. (2009), Nonparametric identification of finite mixture models of dynamic discrete choices, *Econometrica*, 77, 135-175.

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

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

Lu, X. and Su, L. (2016), Determining the number of groups in latent panel structures with an application to income and democracy, *Working Paper,* School of Economics, Singapore Management University.

Mas-Colell, A., Whinston, M. D., and Green, J. R. (1995), *Microeconomic theory*, New York: Oxford university press.

Neumark, D. and Wascher, W. (1992), Employment effects of minimum and submininum wages: panel data on state minimum wage laws, *Industrial and Labor Relations Review*, 46, 55-81.

Neumark, D. and Wascher, W. (2000), Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: Comment, *American Economic Review*, 90, 1362-1396.

Neumark, D. and Wascher, W. (2007), Minimum wages, the earned income tax credit and employment: evidence from the post-welfare reform era, *NBER Working Paper* 12915, NBER.

Park, M. Y., Hastie, T., and Tibshirani, R. (2007), Averaged gene expressions for regression, *Biostatistics*, 8, 212-227.

Phillips, P. C. B. and Sul, D. (2007), Transition modeling and econometric convergence tests, *Econometrica*, 75, 1771-1855.

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

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

Su, L. and Chen, Q. (2013), Testing homogeneity in panel data models with interactive fixed effects, *Econometric Theory*, 29, 1079-1135.

Su, L., Shi, Z., and Phillips, P. C. B. (2016), Identifying latent structures in panel data, *Econometrica,* 84, 2215-2264.

Sun, Y. (2005), Estimation and inference in panel structure models, *Working Paper*, Dept. of Economics, UCSD.

Tibshirani R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

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

Appendix

A Proofs of the Main Results

This appendix provides proofs of the main results in the above paper. Throughout we use M to denote a generic positive constant that may vary across lines. References are made in this Appendix to Lemma B.1, which is a technical result contained in Appendix B, a supplementary document to the present paper.

The proof of Theorem 3.1 makes use of the following lemma.

Lemma A.1 *Suppose that Assumption A1 holds. Then for each* $k = 1, ..., K$,
(i) $P\left(\mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \leq c_1/2\right) = o\left(T^{-1}\right)$,
(ii) $P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i\right\| \geq \frac{M\ln(N_kT)}{\sqrt{N_kT}} + \frac{M[\ln(T)]^2}{T}\right) = o\left(T^{-1}\right)$ *for some* $M > 0$,
(iii) $P\left(\max_{1\leq i\leq N}\mu_{\max}\left(\frac{1}{T}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \geq 2c_2\right) = o\left(T^{-1}\right)$.

Proof of Lemma A.1. (i) First, using $\frac{1}{T}\sum_{t=1}^T \tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}' = \frac{1}{T}\sum_{t=1}^T \mathbf{x}_{it}\mathbf{x}_{it}' - \bar{\mathbf{x}}_i\bar{\mathbf{x}}_i'$ we employ the decomposition

$$\begin{aligned}
\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i &= \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^T \mathbb{E}\left(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right) + \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^T\left[\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}' - \mathbb{E}\left(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)\right] \\
&= \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^T \mathbb{E}\left(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right) + \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^T\left[\mathbf{x}_{it}\mathbf{x}_{it}' - \mathbb{E}\left(\mathbf{x}_{it}\mathbf{x}_{it}'\right)\right] \\
&\quad -\frac{1}{N_k}\sum_{i\in G_k^0}\left[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i' - \mathbb{E}(\bar{\mathbf{x}}_i)\mathbb{E}(\bar{\mathbf{x}}_i')\right] + \frac{1}{N_k}\sum_{i\in G_k^0}\mathrm{Cov}(\bar{\mathbf{x}}_i, \bar{\mathbf{x}}_i).
\end{aligned}$$

It follows that

$$\mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \geq \mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbb{E}\left(\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\right)\right) - \left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left[\mathbf{x}_{it}\mathbf{x}_{it}' - \mathbb{E}\left(\mathbf{x}_{it}\mathbf{x}_{it}'\right)\right]\right\| - \left\|\frac{1}{N_k}\sum_{i\in G_k^0}\left[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i' - \mathbb{E}(\bar{\mathbf{x}}_i)\mathbb{E}(\bar{\mathbf{x}}_i')\right]\right\|.$$

By Lemma B.1(i) of the supplementary document Appendix B, we have

$$P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left[\mathbf{x}_{it}\mathbf{x}_{it}' - \mathbb{E}\left(\mathbf{x}_{it}\mathbf{x}_{it}'\right)\right]\right\| \geq c_1/4\right) = o\left((N_kT)^{-1}\right).$$

Using Lemma B.1(ii), the fact that $\max_{1\leq i\leq N}\|\mathbb{E}(\bar{\mathbf{x}}_i)\| \leq M$ for some $M<\infty$, and the representation $\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i' - \mathbb{E}(\bar{\mathbf{x}}_i)\mathbb{E}(\bar{\mathbf{x}}_i') = \bar{\mathbf{x}}_i\left[\bar{\mathbf{x}}_i - \mathbb{E}(\bar{\mathbf{x}}_i)\right]' + \left[\bar{\mathbf{x}}_i - \mathbb{E}(\bar{\mathbf{x}}_i)\right]\mathbb{E}(\bar{\mathbf{x}}_i')$, we can readily show that

$$P\left(\left\|\frac{1}{N_k}\sum_{i\in G_k^0}\left[\bar{\mathbf{x}}_i\bar{\mathbf{x}}_i' - \mathbb{E}(\bar{\mathbf{x}}_i)\mathbb{E}(\bar{\mathbf{x}}_i')\right]\right\| \geq c_1/4\right) = o\left(T^{-1}\right).$$

It follows that with probability $1-o\left(T^{-1}\right)$ we have $\mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \geq c_1 - c_1/4 - c_1/4 \geq c_1/2$. That is, $P\left(\mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \leq c_1/2\right) = o\left(T^{-1}\right)$.

(ii) We make the following decomposition

$$\begin{aligned}\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i &= \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\left(\mathbf{x}_{it}-\bar{\mathbf{x}}_i\right)\varepsilon_{it} \\ &= \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbf{x}_{it}\varepsilon_{it} - \frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mu_i\varepsilon_{it} - \frac{1}{N_k}\sum_{i\in G_k^0}\left(\bar{\mathbf{x}}_i-\mu_i\right)\bar{\varepsilon}_i,\end{aligned}$$

where $\bar{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T}\varepsilon_{it}$. By Lemma B.1(i), there exists large $M>0$ such that

$$\begin{aligned}P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbf{x}_{it}\varepsilon_{it}\right\| \geq \frac{M\ln\left(N_kT\right)}{2\sqrt{N_kT}}\right) &= o\left((N_kT)^{-1}\right), \text{ and} \\ P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mu_i\varepsilon_{it}\right\| \geq \frac{M\ln\left(N_kT\right)}{2\sqrt{N_kT}}\right) &= o\left((N_kT)^{-1}\right).\end{aligned}$$

By Lemma B.1(ii), $P\left(\max_{i\in G_k^0}\|\bar{\mathbf{x}}_i-\mu_i\| \geq \frac{\sqrt{M}\ln(T)}{\sqrt{T}}\right) = o\left(T^{-1}\right)$ and $P\left(\max_{i\in G_k^0}|\bar{\varepsilon}_i| \geq \frac{\sqrt{M}\ln(T)}{\sqrt{T}}\right) = o\left(T^{-1}\right)$ for some $M>0$. It follows that

$$P\left(\left\|\frac{1}{N_k}\sum_{i\in G_k^0}\left(\bar{\mathbf{x}}_i-\mu_i\right)\bar{\varepsilon}_i\right\| \geq \frac{M\left[\ln\left(T\right)\right]^2}{T}\right) = o\left(T^{-1}\right).$$

Consequently,

$$
\begin{aligned}
& P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i\right\| \geq \frac{M\ln(N_kT)}{\sqrt{N_kT}} + \frac{M[\ln(T)]^2}{T}\right) \\
\leq\ & P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mathbf{x}_{it}\varepsilon_{it}\right\| \geq \frac{M\ln(N_kT)}{2\sqrt{N_kT}}\right) + P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\sum_{t=1}^{T}\mu_i\varepsilon_{it}\right\| \geq \frac{M\ln(N_kT)}{2\sqrt{N_kT}}\right) \\
& + P\left(\left\|\frac{1}{N_k}\sum_{i\in G_k^0}(\bar{\mathbf{x}}_i-\mu_i)\,\bar{\varepsilon}_i\right\| \geq \frac{M[\ln(T)]^2}{T}\right) \\
=\ & o\left(T^{-1}\right).
\end{aligned}
$$

(iii) In view of the fact $\frac{1}{T}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(\mathbf{x}_{it}\mathbf{x}_{it}') + \frac{1}{T}\sum_{t=1}^{T}[\mathbf{x}_{it}\mathbf{x}_{it}' - \mathbb{E}(\mathbf{x}_{it}\mathbf{x}_{it}')] - \bar{\mathbf{x}}_i\bar{\mathbf{x}}_i'$, we have

$$
\mu_{\max}\left(\frac{1}{T}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \leq \mu_{\max}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(\mathbf{x}_{it}\mathbf{x}_{it}'\right)\right) + \left\|\frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}_{it}\mathbf{x}_{it}' - \mathbb{E}\left(\mathbf{x}_{it}\mathbf{x}_{it}'\right)\right]\right\|.
$$

As in the proof of (i), we can readily argue that with probability $1-o\left(T^{-1}\right)$ we have $\mu_{\max}\left(\frac{1}{T}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right) \leq c_2 + c_2 = 2c_2$. This concludes the proof of the lemma. ■

Proof of Theorem 3.1. To prove the theorem, we follow Ke, Fan and Wu (2015, KFW) and prove that with a high probability the Panel-CARDS has a strictly local minimizer given by the oracle estimator $\hat{\boldsymbol{\beta}}^{oracle}$. Recall that $\hat{\boldsymbol{\beta}}^{oracle}$ is obtained with knowledge of the true grouping structure.

First, we introduce the restricted parameter space

$$
M_{\mathcal{G}} = \{\boldsymbol{\beta}\in\mathbb{R}^{Np} : \boldsymbol{\beta}_i = \boldsymbol{\beta}_j \text{ for any } i,j\in G_k^0, 1\leq k\leq K\}. \tag{A.1}
$$

Note that $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_N')'$ and the set $\left\{G_k^0\right\}_{k=1}^K$ denotes the true grouping structure. So $M_{\mathcal{G}}$ is connected with the parameter space of the oracle estimator. We define two mappings:

$$
S : M_{\mathcal{G}} \to \mathbb{R}^{Kp} \text{ and } S^* : \mathbb{R}^{Np} \to \mathbb{R}^{Kp}, \tag{A.2}
$$

where $S(\boldsymbol{\beta})$ is a $Kp\times 1$ vector whose k-th block (the length of a block is p) is the common slope vector $(\boldsymbol{\alpha}_k)$ of group k, and $S^*(\boldsymbol{\beta})$ is a $Kp\times 1$ vector whose k-th block (the length of a block is p) is given by $\frac{1}{N_k}\sum_{i\in G_k^0}\boldsymbol{\beta}_i$, the mean value of slope vectors in group k. Apparently, S and S^* are the same when the domain of S^* is also restricted to be $M_{\mathcal{G}}$. In addition, $\boldsymbol{\alpha}^0 = S(\boldsymbol{\beta}^0)$ and $\hat{\boldsymbol{\alpha}}^{oracle} = S(\hat{\boldsymbol{\beta}}^{oracle})$.

The objective function is $Q_{NT}(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{NT}(\boldsymbol{\beta})$, where $L_{NT}(\boldsymbol{\beta}) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i)^2$ and $P_{NT}(\boldsymbol{\beta}) = P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta})$. For any $\boldsymbol{\alpha}\in\mathbb{R}^{Kp}$, define

$$
\begin{aligned}
L_{NT}^{\mathcal{G}}(\boldsymbol{\alpha}) &= L_{NT}(S^{-1}(\boldsymbol{\alpha})), \quad P_{NT}^{\mathcal{G}}(\boldsymbol{\alpha}) = P_{NT}(S^{-1}(\boldsymbol{\alpha})), \text{ and} \\
Q_{NT}^{\mathcal{G}}(\boldsymbol{\alpha}) &= L_{NT}^{\mathcal{G}}(\boldsymbol{\alpha}) + P_{NT}^{\mathcal{G}}(\boldsymbol{\alpha}).
\end{aligned} \tag{A.3}
$$

We need to show that $\hat{\boldsymbol{\beta}}^{oracle}$ is a strictly local minimizer of Q_{NT} with probability at least $1-\epsilon_0 - o(K/T)$. Let $\mathcal{E}_1$ denote the event that the segmentation $\mathcal{B}$ is admissible with the true parameter $\boldsymbol{\beta}^0$. By the conditions in the theorem, $P(\mathcal{E}_1^c)\leq\epsilon_0$ where, for any event $\mathcal{E}$, $\mathcal{E}^c$ denotes its complement.

Next, we prove that

$$P\left(\|\hat{\boldsymbol{\beta}}^{oracle}-\boldsymbol{\beta}^0\|\leq M\sqrt{K\left(\ln T\right)^2/T}\right)\geq 1-o(K/T) \text{ for some } M>0. \tag{A.4}$$

Define the event $\mathcal{E}_0=\left\{\mu_{\min}\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)>c_1/2\right\}$. Using $\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0=\left(\sum_{i\in G_k^0}\frac{1}{T}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1}\sum_{i\in G_k^0}\frac{1}{T}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i$ and by Lemma A.1, we have uniformly in k

$$\begin{aligned}
& P\left\{\sqrt{N_k}\left\|\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0\right\|\geq M\ln T/\sqrt{T}\right\}\\
= & P\left\{\sqrt{N_k}\left\|\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1}\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i\right\|\geq M\ln T/\sqrt{T}\right\}\\
\leq & P\left\{\sqrt{N_k}\left\|\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1}\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i\right\|\geq M\ln T/\sqrt{T},\mathcal{E}_0\right\}+P\left(\mathcal{E}_0^c\right)\\
\leq & P\left\{\sqrt{N_k}\left\|\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1}\right\|\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i\right\|\geq M\ln T/\sqrt{T},\mathcal{E}_0\right\}+o\left(T^{-1}\right)\\
\leq & P\left(\left\|\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i\right\|\geq\left(\frac{c_1}{2}\right)M\ln T/\sqrt{N_kT}\right)+o\left(T^{-1}\right)=o\left(T^{-1}\right),
\end{aligned}$$

where $P\left(A,B\right)$ denotes $P\left(A\cap B\right)$. With this, we can readily show that

$$\begin{aligned}
P\left(\left\|\hat{\boldsymbol{\beta}}^{oracle}-\boldsymbol{\beta}^0\right\|^2\geq M^2K\left(\ln T\right)^2/T\right) &= P\left(\sum_{k=1}^K N_k\left\|\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0\right\|^2\geq M^2K\left(\ln T\right)^2/T\right)\\
&\leq \sum_{k=1}^K P\left(N_k\left\|\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0\right\|^2\geq M^2\left(\ln T\right)^2/T\right)=o\left(K/T\right).
\end{aligned}$$

Thus (A.4) follows.

Now we consider a small neighborhood of $\boldsymbol{\beta}^0$

$$\mathcal{W}_{NT}^0\equiv\left\{\boldsymbol{\beta}\in\mathbb{R}^{Np}:\|\boldsymbol{\beta}-\boldsymbol{\beta}^0\|<M\ln T\sqrt{K/T}\right\}. \tag{A.5}$$

By (A.4), there exists a set $\mathcal{E}_2$ with $P(\mathcal{E}_2^c)\leq o(K/T)$ and $\|\hat{\boldsymbol{\beta}}^{oracle}-\boldsymbol{\beta}^0\|\leq M\ln T\sqrt{K/T}$ over $\mathcal{E}_2$. For an element $\boldsymbol{\beta}\in\mathcal{W}_{NT}^0$ and $\boldsymbol{\beta}^*=S^{-1}\circ S^*(\boldsymbol{\beta})$. We want to show

(i) Over the set $\mathcal{E}_1\cap\mathcal{E}_2$,

$$Q_{NT}(\boldsymbol{\beta}^*)\geq Q_{NT}(\hat{\boldsymbol{\beta}}^{oracle}) \tag{A.6}$$

and the inequality is strict when $\boldsymbol{\beta}^*\neq\hat{\boldsymbol{\beta}}^{oracle}$.

(ii) There is a set $\mathcal{E}_3$ (to be defined) with $P(\mathcal{E}_3^c)\leq o(T^{-1})$. Over the set $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$, there exists a set $\mathcal{W}_{NT}$ which contains $\hat{\boldsymbol{\beta}}^{oracle}$ such that

$$Q_{NT}(\boldsymbol{\beta})\geq Q_{NT}(\boldsymbol{\beta}^*) \tag{A.7}$$

for any $\boldsymbol{\beta}\in\mathcal{W}_{NT}$, and the inequality is strict when $\boldsymbol{\beta}\neq\boldsymbol{\beta}^*$.

If both (i) and (ii) hold, then we have $Q_{NT}(\boldsymbol{\beta}) \geq Q_{NT}(\hat{\boldsymbol{\beta}}^{oracle})$ for any $\boldsymbol{\beta} \in \mathcal{W}_{NT}$ and $\hat{\boldsymbol{\beta}}^{oracle}$ is a strict local minimizer of Q_{NT} over the set $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$. We prove these two claims in Propositions A.2 and A.3 below. ■

Proposition A.2 *Suppose that the conditions in Theorem 3.1 hold. Then* $Q_{NT}(\boldsymbol{\beta}^*) \geq Q_{NT}(\hat{\boldsymbol{\beta}}^{oracle})$ *on the set* $\mathcal{E}_1 \cap \mathcal{E}_2$ *and the inequality is strict when* $\boldsymbol{\beta}^* \neq \hat{\boldsymbol{\beta}}^{oracle}$.

Proof of Proposition A.2. We demonstrate that

$$P_{NT}^{\mathcal{G}}(S^*(\boldsymbol{\beta})) = \text{Constant for any } \boldsymbol{\beta} \in \mathcal{W}_{NT}^0. \tag{A.8}$$

Recall that $V_{kl} = G_k^0 \cap B_l$ for $k = 1, 2, \cdots, K$ and $l = 1, 2, \cdots, L$. For any $\boldsymbol{\beta} \in \mathcal{W}_{NT}^0$, denote $\boldsymbol{\alpha} = S^*(\boldsymbol{\beta})$. Define $n_{km}^{(1)} = \sum_{l=1}^{L-1}(|V_{kl}||V_{m(l+1)}| + |V_{ml}||V_{k(l+1)}|)$,[10] which is the number of between-segment penalty terms imposed on segments k and m. Similarly, define $n_{km}^{(2)} = 2\sum_{l=1}^{L}|V_{kl}||V_{ml}|$ as the number of within-segment penalty terms. Then

$$P_{NT}^{\mathcal{G}}(\boldsymbol{\alpha}) = \lambda_1 \sum_{1 \leq k < m \leq K} n_{km}^{(1)} \rho_1(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_m\|_1) + \lambda_2 \sum_{1 \leq k < m \leq K} n_{km}^{(2)} \rho_2(\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_m\|_1), \tag{A.9}$$

where $\rho_j(t) = \lambda_j^{-1} p_{\lambda_j}(t)$ for $j = 1, 2$. In view of the fact that

$$\begin{aligned}
\min_{1 \leq k < m \leq K} \|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_m\|_1 &= \min_{1 \leq k < m \leq K} \|\left(\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\right) + \left(\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_m^0\right) - \left(\boldsymbol{\alpha}_m - \boldsymbol{\alpha}_m^0\right)\|_1 \\
&\geq \min_{1 \leq k < m \leq K} \left\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_m^0\right\|_1 - 2 \max_{1 \leq k \leq K} \left\|\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_k^0\right\|_1 \\
&\geq 2b_{NT} - 2p\|\boldsymbol{\beta} - \boldsymbol{\beta}^0\|_\infty \geq 2b_{NT} - 2pM \ln T \sqrt{K/T} > b_{NT} > a \max\{\lambda_1, \lambda_2\}
\end{aligned}$$

by Assumption A3, $P_{NT}^{\mathcal{G}}(\boldsymbol{\alpha})$ in (A.9) is constant on $\mathcal{W}_{NT}^0$ by Assumption A2.

Since $L_{NT}^{\mathcal{G}}(\boldsymbol{\alpha})$ is convex with respect to $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}^{oracle}$ minimizes $L_{NT}^{\mathcal{G}}(\boldsymbol{\alpha})$, we have

$$L_{NT}^{\mathcal{G}}(S^*(\boldsymbol{\beta})) \geq L_{NT}^{\mathcal{G}}\left(\hat{\boldsymbol{\alpha}}^{oracle}\right)$$

for any $\boldsymbol{\alpha} = S^*(\boldsymbol{\beta})$ and the above inequality is strict whenever $S^*(\boldsymbol{\beta}) \neq \hat{\boldsymbol{\alpha}}^{oracle}$, or equivalently, $\boldsymbol{\beta}^* \neq S^{-1}(\hat{\boldsymbol{\alpha}}^{oracle}) = \hat{\boldsymbol{\beta}}^{oracle}$. The conclusion then follows by observing that on $\mathcal{E}_1 \cap \mathcal{E}_2$,

$$\begin{aligned}
Q_{NT}(\boldsymbol{\beta}^*) &= Q_{NT}(S^{-1} \circ S^*(\boldsymbol{\beta})) = Q_{NT}^{\mathcal{G}}(S^*(\boldsymbol{\beta})) = L_{NT}^{\mathcal{G}}(S^*(\boldsymbol{\beta})) + P_{NT}^{\mathcal{G}}(S^*(\boldsymbol{\beta})) \\
&= L_{NT}^{\mathcal{G}}(S^*(\boldsymbol{\beta})) + \text{Constant}
\end{aligned}$$

and, similarly, $Q_{NT}(\hat{\boldsymbol{\beta}}^{oracle}) = L_{NT}^{\mathcal{G}}(\hat{\boldsymbol{\alpha}}^{oracle}) + P_{NT}^{\mathcal{G}}(S^*(\hat{\boldsymbol{\alpha}}^{oracle})) = L_{NT}^{\mathcal{G}}(\hat{\boldsymbol{\alpha}}^{oracle}) + \text{Constant}$. ■

Proposition A.3 *Suppose that the conditions in Theorem 3.1 hold. Then there exists a set* $\mathcal{W}_{NT}$ *which contains* $\hat{\boldsymbol{\beta}}^{oracle}$ *such that* $Q_{NT}(\boldsymbol{\beta}) \geq Q_{NT}(\boldsymbol{\beta}^*)$ *on the set* $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ *for any* $\boldsymbol{\beta} \in \mathcal{W}_{NT}$, *and the inequality is strict when* $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$.

[10] Since the ordered segmentation is admissible, we note here that many of the $V_{kl}^{'}$s are empty with cardinality 0.

Proof of Proposition A.3. We construct a subset of $\mathcal{W}_{NT}^0$ defined by

$$\mathcal{W}_{NT} = \mathcal{W}_{NT}^0 \cap \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oracle}\| \leq t_{NT}\}, \tag{A.10}$$

where t_{NT} is a positive sequence such that $\frac{t_{NT}}{NN_{\min}} \ll \lambda_2$ and $t_{NT} \ll \lambda_1$. Recall that $\boldsymbol{\beta}^* = S^{-1} \circ S^*(\boldsymbol{\beta})$, which implies $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|$ for any $\boldsymbol{\beta}' \in \mathcal{M}_{\mathcal{G}}$. In particular, we have $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oracle}\|$. Consequently, it suffices to prove the proposition by showing (A.7) holds for any $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| \leq t_{NT}$, and the inequality is strict when $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$.

We now analyze how $Q_{NT}(\boldsymbol{\beta})$ responds to the change of $\boldsymbol{\beta} \in \mathcal{W}_{NT}$. We make the following decomposition

$$Q_{NT}(\boldsymbol{\beta}) - Q_{NT}(\boldsymbol{\beta}^*) = [L_{NT}(\boldsymbol{\beta}) - L_{NT}(\boldsymbol{\beta}^*)] + [P_{NT}(\boldsymbol{\beta}) - P_{NT}(\boldsymbol{\beta}^*)] \equiv I_1 + I_2, \text{ say.} \tag{A.11}$$

The basic idea is to demonstrate that upon moving from $\boldsymbol{\beta}$ to $\boldsymbol{\alpha} = S^*(\boldsymbol{\beta})$, the decrease in the penalty term I_2 dominates the increase in the least squares function I_1 with high probability. By the Cauchy-Schwarz inequality, $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2 \leq \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1^2 \leq p\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_2^2$. For I_2 we have

$$\begin{aligned}
I_2 &= P_{NT}(\boldsymbol{\beta}) - P_{NT}(\boldsymbol{\beta}^*) \\
&= \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}} p_{\lambda_1}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1) + \sum_{l=1}^{L} \sum_{i \in B_l, j \in B_l} p_{\lambda_2}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1) \\
&\quad - \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}} p_{\lambda_1}(\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^*\|_1) - \sum_{l=1}^{L} \sum_{i \in B_l, j \in B_l} p_{\lambda_2}(\|\boldsymbol{\beta}_i^* - \boldsymbol{\beta}_j^*\|_1) \\
&= \lambda_1 \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}, i \overset{\mathcal{G}}{\sim} j} \rho_1(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1) + \lambda_2 \sum_{l=1}^{L} \sum_{i \in B_l, j \in B_l, i \overset{\mathcal{G}}{\sim} j} \rho_2(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1) \\
&\geq \lambda_1 \sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}, i \overset{\mathcal{G}}{\sim} j} \rho_1'\left(\frac{2\sqrt{p}t_{NT}}{\sqrt{N_{\min}}}\right) \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1 + \lambda_2 \sum_{l=1}^{L} \sum_{i \in B_l, j \in B_l, i \overset{\mathcal{G}}{\sim} j} \rho_2'\left(\frac{2\sqrt{p}t_{NT}}{\sqrt{N_{\min}}}\right) \|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1,
\end{aligned} \tag{A.12}$$

where $i \overset{\mathcal{G}}{\sim} j$ means i and j are in the same true group in which case $\boldsymbol{\beta}_i^* = \boldsymbol{\beta}_j^*$, the third equality follows from the proof of (A.8), and the last inequality follow from the concavity of $\rho_1(\cdot)$ and $\rho_2(\cdot)$ and for i, j in the same true group, $\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1 \leq 2\sqrt{p}\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|/\sqrt{N_{\min}} \leq 2\sqrt{p}t_{NT}/\sqrt{N_{\min}}$.

For I_1, we apply a Taylor development, giving

$$\begin{aligned}
I_1 &= L_{NT}(\boldsymbol{\beta}) - L_{NT}(\boldsymbol{\beta}^*) \\
&= \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i)^2 - \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} \sum_{t=1}^{T} (\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_k^*)^2 \\
&= \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} (\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_i)'(\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_i) - \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} (\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_k^*)'(\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\boldsymbol{\beta}_k^*) \\
&= -\frac{1}{NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} (\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\breve{\boldsymbol{\beta}}_{k(i)})'\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_i - \boldsymbol{\beta}_k^*) \\
&= -\frac{1}{NT} \sum_{k=1}^{K} \sum_{l=d_k}^{u_k} \sum_{i \in V_{kl}} (\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\breve{\boldsymbol{\beta}}_{k(i)})'\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_i - \boldsymbol{\beta}_k^*),
\end{aligned} \tag{A.13}$$

where $\breve{\boldsymbol{\beta}}_{k(i)}$ denotes the intermediate value that lies between $\boldsymbol{\beta}_i$ and $\boldsymbol{\beta}_k^*$ elementwise. Let $\mathbf{z}_i = \tilde{\mathbf{x}}_i'(\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\breve{\boldsymbol{\beta}}_{k(i)})$. Noting that $\boldsymbol{\beta}_k^* = \frac{1}{N_k}\sum_{i'\in G_k^0}\boldsymbol{\beta}_{i'} = \frac{1}{N_k}\sum_{l'=d_k}^{u_k}\sum_{i'\in V_{kl'}}\boldsymbol{\beta}_{i'}$, we have

$$\begin{aligned}
I_1 &= -\frac{1}{NT}\sum_{k=1}^{K}\sum_{l=d_k}^{u_k}\sum_{i\in V_{kl}}\mathbf{z}_i'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_k^*) = -\frac{1}{NT}\sum_{k=1}^{K}\sum_{l=d_k}^{u_k}\sum_{i\in V_{kl}}\mathbf{z}_i'\frac{1}{N_k}\sum_{l'=d_k}^{u_k}\sum_{i'\in V_{kl'}}(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}) \\
&= -\frac{1}{2NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{l=d_k}^{u_k}\sum_{l'=d_k}^{u_k}\sum_{i\in V_{kl}}\sum_{i'\in V_{kl'}}(\mathbf{z}_i - \mathbf{z}_{i'})'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}) \\
&= -\frac{1}{2NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{l=d_k}^{u_k}\sum_{i\in V_{kl}}\sum_{i'\in V_{kl}}(\mathbf{z}_i - \mathbf{z}_{i'})'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}) \\
&\quad -\frac{1}{NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{d_k\le l<l'\le u_k}\sum_{i\in V_{kl}}\sum_{i'\in V_{kl'}}(\mathbf{z}_i - \mathbf{z}_{i'})'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}) \\
&= : I_{11} + I_{12}. \qquad (A.14)
\end{aligned}$$

We will evaluate I_{11} and I_{12} in turn. First we transform I_{11} for comparison,

$$\begin{aligned}
I_{11} &= -\frac{1}{2NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{l=d_k}^{u_k}\sum_{i\in V_{kl}}\sum_{i'\in V_{kl}}(\mathbf{z}_i - \mathbf{z}_{i'})'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}) \\
&= -\frac{1}{2NT}\sum_{l=1}^{L}\sum_{k=a_l}^{b_l}\sum_{i\in V_{kl}}\sum_{i'\in V_{kl}}\frac{1}{N_k}(\mathbf{z}_i - \mathbf{z}_{i'})'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}) \\
&= -\frac{1}{NT}\sum_{l=1}^{L}\sum_{i,i'\in B_l, i\overset{\mathcal{G}}{\sim}i'}\boldsymbol{\theta}_{ii'}(\mathbf{z})'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}), \qquad (A.15)
\end{aligned}$$

where $\mathbf{z} = (\mathbf{z}_\mathbf{1}', \cdots, \mathbf{z}_\mathbf{N}')'$, $\boldsymbol{\theta}_{ii'}(\mathbf{z}) = \frac{1}{2N_k}(\mathbf{z}_i - \mathbf{z}_{i'})$, and as before $i \overset{\mathcal{G}}{\sim} i'$ means that i and i' belong to the same true group. Now we change I_{12} to a form that can be easily compared with I_2. By the property of the partition $\mathcal{B}$, we can write

$$\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'} = \frac{1}{\prod_{h=l+1}^{l'-1}|V_{kh}|}\sum_{\{(i_l, i_{l+1}, \cdots, i_{l'}):\ i_l = i, i_{l'} = i'; i_h\in V_{kh}, h = l+1, \cdots, l'-1\}}\ \sum_{h=l}^{l'-1}(\boldsymbol{\beta}_{i_h} - \boldsymbol{\beta}_{i_{h+1}}),$$

where the second summation is a telescope summation with common value $\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i'}$, the first summation is over all possible paths from all sets V_{kh} between V_{kl} and $V_{kl'}$, and the total number of different paths is given by $\prod_{h=l+1}^{l'-1}|V_{kh}|$. For notation consistency, when $l = l' - 1$, we define

$\prod_{h=l+1}^{l'-1}|V_{kh}|=1$. Plugging the expression into I_{12}, we have

$$\begin{aligned}
I_{12} &= -\frac{1}{NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{d_k\le l<l'\le u_k}\sum_{i\in V_{kl}}\sum_{i'\in V_{kl'}}(\mathbf{z}_i-\mathbf{z}_{i'})'(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}) \\
&= -\frac{1}{NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{d_k\le l<l'\le u_k}\sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_h\in V_{kh},h=l,\cdots,l'\}}\frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{h=l+1}^{l'-1}|V_{kh}|}\sum_{h=l}^{l'-1}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) \\
&= -\frac{1}{NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{d_k\le l<l'\le u_k}S_{ll',k},
\end{aligned}$$

where

$$S_{ll',k}=\sum_{h=l}^{l'-1}\sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}}\frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}).$$

To simplify the last expression, we discuss four cases: (a) $l=h=l'-1$, (b) $l=h<l'-1$, (c) $l<h<l'-1$, and (d) $l<h=l'-1$, and write

$$S_{ll',k}=S_{ll',k}(a)+S_{ll',k}(b)+S_{ll',k}(c)+S_{ll',k}(d),$$

where, for example, $S_{ll',k}(a)$ denotes the summation in $S_{ll',k}$ for which h is restricted to satisfy the conditions in (a). In case (a), we have

$$\begin{aligned}
S_{ll',k}(a) &= \sum_{h=l}^{l'-1}\sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}}\frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}})\mathbf{1}\left\{l=h=l'-1\right\} \\
&= \sum_{i_h\in V_{kh},i_{h+1}\in V_{k,h+1}}\left(\mathbf{z}_{i_l}-\mathbf{z}_{i_{h+1}}\right)'(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) \\
&= \sum_{i\in V_{kh}}\sum_{i'\in V_{k,h+1}}(\mathbf{z}_i-\mathbf{z}_{i'})'(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}).
\end{aligned}$$

In case (b),

$$\begin{aligned}
S_{ll',k}(b) &= \sum_{h=l}^{l'-1}\sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}}\frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}})\mathbf{1}\left\{l=h<l'-1\right\} \\
&= \sum_{i_h\in V_{kh}}\sum_{i_{h+1}\in V_{k,h+1}}\sum_{i_{l'}\in V_{kl'}}\frac{\mathbf{z}'_{i_h}-\mathbf{z}'_{i_{l'}}}{|V_{k,h+1}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) \\
&= \sum_{i_h\in V_{kh}}\sum_{i_{h+1}\in V_{k,h+1}}\frac{|V_{kl'}|}{|V_{k,h+1}|}\left(\mathbf{z}'_{i_h}-\overline{\mathbf{z}}'_{kl'}\right)(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) \\
&= \sum_{i\in V_{kh}}\sum_{i'\in V_{k,h+1}}\frac{|V_{kl'}|}{|V_{k,l+1}|}(\mathbf{z}_i-\overline{\mathbf{z}}_{kl'})'(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}),
\end{aligned}$$

where $\bar{\mathbf{z}}_{kl'} = \frac{1}{|V_{kl'}|}\sum_{j\in V_{kl'}} \mathbf{z}_j$. Similarly, in case (d) we have

$$\begin{aligned} S_{ll',k}(d) &= \sum_{h=l}^{l'-1} \sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}} \frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}})\mathbf{1}\left\{l<h=l'-1\right\} \\ &= \sum_{i\in V_{kh}}\sum_{i'\in V_{k,h+1}} \frac{|V_{kl}|}{|V_{k,l'-1}|}(\bar{\mathbf{z}}_{kl}-\mathbf{z}_{i'})'(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}). \end{aligned}$$

In case (c)

$$\begin{aligned} S_{ll',k}(c) &= \sum_{h=l}^{l'-1} \sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}} \frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}})\mathbf{1}\left\{l<h<l'-1\right\} \\ &= \sum_{h=l+1}^{l'-2} \sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}} \frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) \\ &= \sum_{h=l+1}^{l'-2} \sum_{i_h\in V_{kh},i_{h+1}\in V_{k,h+1}} \sum_{i_l\in V_{kl},i_{l'}\in V_{kl'}} \frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{|V_{kh}||V_{k,h+1}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) \\ &= \sum_{h=l+1}^{l'-2} \sum_{i_h\in V_{kh},i_{h+1}\in V_{k,h+1}} \frac{|V_{kl}||V_{kl'}|}{|V_{kh}||V_{k,h+1}|}(\bar{\mathbf{z}}_{kl}-\bar{\mathbf{z}}_{kl'})'(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}). \end{aligned}$$

It follows that

$$S_{ll',k} = \sum_{h=l}^{l'-1} \sum_{\{(i_l,i_{l+1},\cdots,i_{l'}):i_j\in V_{kj},j=l,\cdots,l'\}} \frac{\mathbf{z}'_{i_l}-\mathbf{z}'_{i_{l'}}}{\prod_{j=l+1}^{l'-1}|V_{kj}|}(\boldsymbol{\beta}_{i_h}-\boldsymbol{\beta}_{i_{h+1}}) = \sum_{h=l}^{l'-1}\sum_{i\in V_{kh}}\sum_{i'\in V_{k(h+1)}} \boldsymbol{\omega}'_{ii',ll',h}(\mathbf{z})(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}),$$

where

$$\boldsymbol{\omega}_{ii',ll',h}(\mathbf{z}) = \begin{cases} \mathbf{z}_i-\mathbf{z}_{i'}, & l=h=l'-1 \\ \frac{|V_{kl'}|}{|V_{k(l+1)}|}(\mathbf{z}_i-\bar{\mathbf{z}}_{kl'}), & l=h<l'-1 \\ \frac{|V_{kl}||V_{kl'}|}{|V_{kh}||V_{k(h+1)}|}(\bar{\mathbf{z}}_{kl}-\bar{\mathbf{z}}_{kl'}), & l<h<l'-1 \\ \frac{|V_{kl}|}{|V_{k(l'-1)}|}(\bar{\mathbf{z}}_{kl}-\mathbf{z}_{i'}), & l<h=l'-1 \end{cases}. \tag{A.16}$$

Then

$$\begin{aligned} I_{12} &= -\frac{1}{NT}\sum_{k=1}^{K}\frac{1}{N_k}\sum_{d_k\le l<l'\le u_k}\sum_{h=l}^{l'-1}\sum_{i\in V_{kh}}\sum_{i'\in V_{k(h+1)}} \boldsymbol{\omega}'_{ii',ll',h}(\mathbf{z})(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}) \\ &= -\frac{1}{NT}\sum_{h=1}^{L-1}\sum_{k=a_h}^{b_h}\sum_{i\in V_{kh},i'\in V_{k(h+1)}}\left[\frac{1}{N_k}\sum_{l=d_k}^{h}\sum_{l'=h+1}^{u_k}\boldsymbol{\omega}'_{ii',ll',h}(\mathbf{z})\right](\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}) \\ &= -\frac{1}{NT}\sum_{h=1}^{L-1}\sum_{i\in B_h,i'\in B_{h+1},i\overset{\mathcal{G}}{\sim}i'}\boldsymbol{\tau}'_{ii'}(\mathbf{z})(\boldsymbol{\beta}_i-\boldsymbol{\beta}_{i'}), \end{aligned} \tag{A.17}$$

where $\boldsymbol{\tau}_{ii'}(\mathbf{z}) = \frac{1}{N_k}\sum_{l=d_k}^{h}\sum_{l'=h+1}^{u_k}\boldsymbol{\omega}_{ii',ll',h}(\mathbf{z})$. Let $G_{kh}^1 = \bigcup_{l\le h} V_{kl}$ and $G_{kh}^2 = \bigcup_{l>h} V_{kl}$. Then by (A.16)

$$\begin{aligned}
\boldsymbol{\tau}_{ii'}(\mathbf{z}) &= \frac{1}{N_k}\sum_{l=d_k}^{h}\sum_{l'=h+1}^{u_k}\boldsymbol{\omega}_{ii',ll',h}(\mathbf{z}) \\
&= \frac{1}{N_k}\sum_{l=d_k}^{h-1}\sum_{l'=h+2}^{u_k}\frac{|V_{kl}||V_{kl'}|}{|V_{kh}||V_{k(h+1)}|}(\bar{\mathbf{z}}_{kl}-\bar{\mathbf{z}}_{kl'}) + \frac{1}{N_k}\sum_{l=d_k}^{h-1}\frac{|V_{kl}|}{|V_{kh}|}(\bar{\mathbf{z}}_{kl}-\mathbf{z}_{i'}) \\
&\quad + \frac{1}{N_k}\sum_{l'=h+2}^{u_k}\frac{|V_{kl'}|}{|V_{k(h+1)}|}(\mathbf{z}_i-\bar{\mathbf{z}}_{kl'}) + \frac{1}{N_k}(\mathbf{z}_i-\mathbf{z}_{i'}) \\
&= \frac{1}{N_k}\sum_{l=d_k}^{h-1}\frac{|V_{kl}|(\sum_{l'=h+1}^{u_k}|V_{kl'}|)}{|V_{kh}||V_{k(h+1)}|}\bar{\mathbf{z}}_{kl} + \frac{1}{N_k}\frac{\sum_{l'=h+1}^{u_k}|V_{kl'}|}{|V_{k(h+1)}|}\mathbf{z}_i \\
&\quad - \frac{1}{N_k}\sum_{l'=h+2}^{u_k}\frac{(\sum_{l=d_k}^{h}|V_{kl}|)|V_{kl'}|}{|V_{kh}||V_{k(h+1)}|}\bar{\mathbf{z}}_{kl'} - \frac{1}{N_k}\frac{\sum_{l=d_k}^{h}|V_{kl}|}{|V_{kh}|}\mathbf{z}_{i'} \\
&= \frac{1}{|V_{kh}||V_{k(h+1)}|}\left(\frac{|G_{kh}^2|}{N_k}\sum_{j\in G_{k(h-1)}^1}\mathbf{z}_j - \frac{|G_{kh}^1|}{N_k}\sum_{j\in G_{k(h+1)}^2}\mathbf{z}_j\right) \\
&\quad + \left(\frac{|G_{kh}^2|}{N_k|V_{k(h+1)}|}\mathbf{z}_i - \frac{|G_{kh}^1|}{N_k|V_{kh}|}\mathbf{z}_{i'}\right). \qquad \text{(A.18)}
\end{aligned}$$

By (A.14), (A.15) and (A.17), we have

$$\begin{aligned}
|I_1| &\le |I_{11}| + |I_{12}| \\
&\le \frac{1}{NT}\sum_{l=1}^{L}\sum_{i,j\in B_l, i\overset{\mathcal{G}}{\sim}j}\|\boldsymbol{\theta}_{ij}(\mathbf{z})\|_1\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_1 + \frac{1}{NT}\sum_{l=1}^{L-1}\sum_{i\in B_l, j\in B_{l+1}, i\overset{\mathcal{G}}{\sim}j}\|\boldsymbol{\tau}_{ij}(\mathbf{z})\|_1\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_1. \qquad \text{(A.19)}
\end{aligned}$$

By (A.11), (A.13) and (A.19), we have

$$\begin{aligned}
Q_{NT}(\boldsymbol{\beta}) - Q_{NT}(\boldsymbol{\beta}^*) &\ge \sum_{l=1}^{L}\sum_{i,j\in B_l, i\overset{\mathcal{G}}{\sim}j}\left[\lambda_2\rho_2'\left(\frac{2\sqrt{p}t_{NT}}{\sqrt{N_{\min}}}\right) - \frac{1}{NT}\|\boldsymbol{\theta}_{ij}(\mathbf{z})\|_1\right]\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_1 \\
&\quad + \sum_{l=1}^{L-1}\sum_{i\in B_l, j\in B_{l+1}, i\overset{\mathcal{G}}{\sim}j}\left[\lambda_1\rho_1'\left(\frac{2\sqrt{p}t_{NT}}{\sqrt{N_{\min}}}\right) - \frac{1}{NT}\|\boldsymbol{\tau}_{ij}(\mathbf{z})\|_1\right]\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_1. \\
&\equiv J_{11} + J_{12} \qquad \text{(A.20)}
\end{aligned}$$

Now we only need to find a high probability event $\mathcal{E}_3$ over which the right hand side of (A.20) is nonnegative, and $P(\mathcal{E}_3)$ should be at least $1-o(T^{-1})$. Noting that

$$\begin{aligned}
\mathbf{z}_i &= \tilde{\mathbf{x}}_i'(\tilde{\mathbf{y}}_i - \tilde{\mathbf{x}}_i\breve{\boldsymbol{\beta}}_{k(i)}) = \tilde{\mathbf{x}}_i'(\tilde{\boldsymbol{\varepsilon}}_i + \tilde{\mathbf{x}}_i\boldsymbol{\beta}_k^0) - \tilde{\mathbf{x}}_i\breve{\boldsymbol{\beta}}_{k(i)}) \\
&= \tilde{\mathbf{x}}_i'\tilde{\boldsymbol{\varepsilon}}_i - \tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^0) - \tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i(\breve{\boldsymbol{\beta}}_{k(i)} - \boldsymbol{\beta}_k^*), \qquad \text{(A.21)}
\end{aligned}$$

we have

$$\begin{aligned}\boldsymbol{\theta}_{ij}(\mathbf{z}) &= \frac{1}{2N_k}\left(\tilde{\mathbf{x}}_i'\tilde{\boldsymbol{\varepsilon}}_i - \tilde{\mathbf{x}}_j'\tilde{\boldsymbol{\varepsilon}}_j\right) - \frac{1}{2N_k}\left(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j\right)\left(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^0\right) \\ &\quad -\frac{1}{2N_k}\left[\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i(\breve{\boldsymbol{\beta}}_{k(i)} - \boldsymbol{\beta}_k^*) - \tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j(\breve{\boldsymbol{\beta}}_{k(j)} - \boldsymbol{\beta}_k^*)\right] \\ &\equiv \boldsymbol{\theta}_{ij,1} - \boldsymbol{\theta}_{ij,2} - \boldsymbol{\theta}_{ij,3}, \text{ say.}\end{aligned}$$

Note that $\boldsymbol{\theta}_{ij,1} = \frac{1}{2N_k}\sum_{t=1}^T\left(\tilde{\mathbf{x}}_{it}\varepsilon_{it} - \tilde{\mathbf{x}}_{jt}\varepsilon_{jt}\right) = \frac{1}{2N_k}\sum_{t=1}^T\left(\mathbf{x}_{it}\varepsilon_{it} - \mathbf{x}_{jt}\varepsilon_{jt}\right) + \frac{1}{2N_k}\left(\bar{\mathbf{x}}_i\bar{\varepsilon}_i - \bar{\mathbf{x}}_j\bar{\varepsilon}_j\right)$. By Lemma B.1, we can readily show that

$$P\left(\max_{1\le k\le K}\max_{i,j\in G_k^0}\frac{1}{T}\left\|\boldsymbol{\theta}_{ij,1}\right\|_1 \ge \frac{M\ln T}{N_{\min}\sqrt{T}}\right) = o\left(T^{-1}\right) \text{ for some } M>0.$$

For $\boldsymbol{\theta}_{ij,2}$, we have by Lemma A.1(iii), with probability $1-o\left(T^{-1}\right)$

$$\begin{aligned}\max_{1\le k\le K}\max_{i,j\in G_k^0}\frac{1}{T}\left\|\boldsymbol{\theta}_{ij,2}\right\|_1 &\le \max_{1\le k\le K}\max_{i,j\in G_k^0}\frac{\sqrt{p}}{2TN_k}\left\|\left(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j'\tilde{\mathbf{x}}_j\right)\left(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^0\right)\right\| \\ &\le \max_{1\le k\le K}\frac{\sqrt{p}}{N_k}\max_{1\le i\le N}\mu_{\max}\left(\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i/T\right)\max_{1\le k\le K}\left\|\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^0\right\| \le \frac{2c_2\sqrt{p}}{N_{\min}}t_{NT}.\end{aligned}$$

Similarly, $\max_{1\le k\le K}\max_{i,j\in G_k^0}\frac{1}{T}\left\|\boldsymbol{\theta}_{ij,3}\right\|_1 \le \frac{2c_2\sqrt{p}}{N_{\min}}t_{NT}$ with probability $1-o\left(T^{-1}\right)$. It follows that with probability $1-o\left(T^{-1}\right)$ we have

$$\begin{aligned}\frac{1}{NT}\max_{1\le k\le K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\theta}_{ij}(\mathbf{z})\right\|_1 &\le \frac{1}{NT}\max_{1\le k\le K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\theta}_{ij,1} - \boldsymbol{\theta}_{ij,2} - \boldsymbol{\theta}_{ij,3}\right\|_1 \\ &\le \frac{M\ln T}{NN_{\min}\sqrt{T}} + \frac{4c_2\sqrt{p}}{NN_{\min}}t_{NT} \le \frac{M}{NN_{\min}}\left(\frac{\ln T}{\sqrt{T}} + t_{NT}\right).\end{aligned}$$

Define

$$\mathcal{E}_{31} = \left\{\frac{1}{NT}\max_{1\le k\le K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\theta}_{ij}(\mathbf{z})\right\|_1 \le \frac{M}{NN_{\min}}\left(\frac{\ln T}{\sqrt{T}} + t_{NT}\right)\right\}.$$

By choosing sufficiently small t_{NT}, we have $\frac{1}{NN_{\min}}\left(\frac{\ln T}{\sqrt{T}} + t_{NT}\right) \ll \lambda_2$. It follows that $J_{11} > 0$ over the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_{31}$ with $P\left(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_{32}\right) = 1 - o\left(T^{-1}\right)$.

Next, we consider J_{12}. By the linearity of $\tau_{ii'}(\cdot)$ and (A.21), we can write

$$\boldsymbol{\tau}_{ii'}(\mathbf{z}) = \boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}'\boldsymbol{\varepsilon}) - \boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}_{(1)}) - \boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}_{(2)}),$$

where $\tilde{\mathbf{X}}$ denotes an $NT\times Np$ block diagonal matrix with the ith diagonal block given by $\tilde{\mathbf{x}}_i$, $\tilde{\mathbf{X}}_{(1)}$ is $Np\times 1$ vector with typical block $\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i(\boldsymbol{\beta}_k^* - \boldsymbol{\beta}_k^0)$ for $i\in G_k^0$, and $\tilde{\mathbf{X}}_{(2)}$ is $Np\times 1$ vector with typical block $\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i(\breve{\boldsymbol{\beta}}_{k(i)} - \boldsymbol{\beta}_k^*)$ for $i\in G_k^0$. By (A.18),

$$\begin{aligned}\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}'\boldsymbol{\varepsilon}) &= \frac{1}{|V_{kh}||V_{k(h+1)}|}\left(\frac{|G_{kh}^2|}{N_k}\sum_{j\in G_{k(h-1)}^1}\tilde{\mathbf{x}}_j'\boldsymbol{\varepsilon}_j - \frac{|G_{kh}^1|}{N_k}\sum_{j\in G_{k(h+1)}^2}\tilde{\mathbf{x}}_j'\boldsymbol{\varepsilon}_j\right) \\ &\quad + \left(\frac{|G_{kh}^2|}{N_k|V_{k(h+1)}|}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i - \frac{|G_{kh}^1|}{N_k|V_{kh}|}\tilde{\mathbf{x}}_{i'}'\boldsymbol{\varepsilon}_{i'}\right).\end{aligned}$$

By Lemma B.1, we can readily show that with probability $1-o\left(T^{-1}\right)$ we have

$$\frac{1}{T}\left\|\sum_{j\in G^1_{k(h-1)}}\tilde{\mathbf{x}}_j'\boldsymbol{\varepsilon}_j\right\|_1 \leq \frac{M\ln T\sqrt{\left|G^1_{k(h-1)}\right|}}{T^{1/2}} \text{ and } \frac{1}{T}\left\|\sum_{j\in G^2_{k(h+1)}}\tilde{\mathbf{x}}_j'\boldsymbol{\varepsilon}_j\right\|_1 \leq \frac{M\ln T\sqrt{\left|G^2_{k(h+1)}\right|}}{T^{1/2}}$$

It follows that with probability $1-o\left(T^{-1}\right)$,

$$\frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G^0_k}\left\|\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}'\boldsymbol{\varepsilon})\right\|_1 \leq \frac{M\ln T}{NT^{1/2}}\max_{k,h}\mathbb{S}_{kh},$$

where $(\mathbb{S}_{kh})^2 = \frac{4}{|V_{kh}|^2|V_{k(h+1)}|^2}\frac{|G^2_{kh}|^2|G^1_{k(h-1)}|+|G^1_{kh}|^2|G^2_{k(h+1)}|}{N_k^2} + \frac{4|G^2_{kh}|^2}{N_k^2|V_{k(h+1)}|^2}+\frac{4|G^1_{kh}|^2}{N_k^2|V_{kh}|^2}$. Below we use the fact that

$$|G^1_{k(h-1)}| < \left|G^1_{kh}\right| \leq N_k,\ |G^2_{k(h+1)}| < \left|G^2_{kh}\right| \leq N_k, \text{ and } \left|G^1_{kh}\right| + \left|G^2_{kh}\right| = N_k.$$

We consider four subcases: (1) $h > d_k$, $h+1 < u_k$, (2) $h > d_k, h+1 = u_k$, (3) $h = d_k, h+1 < u_k$, and (4) $h = d_k$, $h+1 = u_k$. In subcase (1), we have $|V_{kh}| = |B_h|$, $|V_{k(h+1)}| = |B_{h+1}|$, and

$$(\mathbb{S}_{kh})^2 \leq \frac{4N_k}{|B_h|^2|B_{h+1}|^2} + \frac{4}{|B_{h+1}|^2} + \frac{4}{|B_h|^2}.$$

In subcase (2), we have $|V_{kh}| = |B_h|$, $|G^2_{kh}| = \left|V_{k(h+1)}\right|$, and

$$(\mathbb{S}_{kh})^2 \leq \frac{4N_k}{|B_h|^2|V_{k(h+1)}|^2} + \frac{4}{N_k^2}+\frac{4}{|B_h|^2}.$$

In subcase (3) we have $|G^1_{kh}| = |V_{kh}|$, $|V_{k(h+1)}| = |B_{h+1}|$, and

$$(\mathbb{S}_{kh})^2 \leq \frac{4N_k}{|V_{kh}|^2|B_{h+1}|^2} + \frac{4}{|B_{h+1}|^2}+\frac{4}{N_k^2}.$$

In subcase (4), we have $|G^1_{kh}| = |V_{kh}|$, $|G^2_{kh}| = \left|V_{k(h+1)}\right|$, and

$$(\mathbb{S}_{kh})^2 \leq \frac{8}{N_k^2}.$$

In sum, $(\mathbb{S}_{kh})^2 \leq \frac{12N_k}{\min\{N_k^3,\ \min_{d_k\leq l\leq u_k}|B_l|^2\}} =: 12\phi_k$. It follows that with probability $1-o\left(T^{-1}\right)$

$$\frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G^0_k}\left\|\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}'\boldsymbol{\varepsilon})\right\|_1 \leq \frac{M\ln T}{NT^{1/2}}\sqrt{\phi_k}.$$

By the same token, we can show that with probability $1-o\left(T^{-1}\right)$

$$\frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G^0_k}\left\|\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}_{(s)})\right\|_1 \leq \frac{M\sqrt{\phi_k}}{N}\max_{1\leq k\leq K}\left\|\boldsymbol{\beta}^*_k-\boldsymbol{\beta}^0_k\right\| \leq \frac{M\sqrt{\phi_k}}{N}t_{NT} \text{ for } s=1,2.$$

Then with probability $1-o\left(T^{-1}\right)$ we have

$$\begin{aligned}\frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\tau}_{ij}(\mathbf{z})\right\|_1 &= \frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}'\boldsymbol{\varepsilon})-\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}_{(1)})-\boldsymbol{\tau}_{ii'}(\tilde{\mathbf{X}}_{(2)})\right\|_1 \\ &\leq \frac{M}{N}\left(\frac{\ln T}{T^{1/2}}+t_{NT}\right)\sqrt{\max_{1\leq k\leq K}\phi_k}.\end{aligned}$$

Define

$$\mathcal{E}_{32}=\left\{\frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\tau}_{ij}(\mathbf{z})\right\|_1\leq\frac{M}{N}\left(\frac{\ln T}{T^{1/2}}+t_{NT}\right)\sqrt{\max_{1\leq k\leq K}\phi_k}\right\}.$$

By choosing sufficiently small t_{NT} (e.g., $t_{NT}=M\ln T/T^{1/2}$), we have $\frac{1}{N}\left(\frac{\ln T}{T^{1/2}}+t_{NT}\right)\sqrt{\phi_k}\ll\lambda_1$. By the conditions on λ_1, λ_2, and ϕ_k, we have $J_{12}>0$ on the event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_{32}$ with $P\left(\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_{32}\right)=1-o\left(T^{-1}\right)$.

In sum, over the event $\mathcal{E}_1\cap\mathcal{E}_2\cap\mathcal{E}_3$ with $\mathcal{E}_3=\mathcal{E}_{31}\cap\mathcal{E}_{32}$, we have $Q_{NT}(\boldsymbol{\beta})\geq Q_{NT}(\boldsymbol{\beta}^*)$ for any $\boldsymbol{\beta}\in\mathcal{W}_{NT}$ and the strict inequality holds for $\boldsymbol{\beta}\neq\boldsymbol{\beta}^*$. ■

Proof of Theorem 3.2. (i) By Theorem 3.1, $P(\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{oracle})\to 1$ provided $\epsilon_0\equiv\epsilon_{0T}\to 0$ as $T\to\infty$. It follows that $P(\hat{K}=K)\to 1$ and $P(\hat{G}_1=G_1^0,...,\hat{G}_K=G_K^0|\hat{K}=K)\to 1$ as $T\to\infty$, perhaps after suitable relabeling among the G_k^0's. In addition,

$$P\left(\hat{G}_1=G_1^0,...,\hat{G}_K=G_K^0\right)=P\left(\hat{G}_1=G_1^0,...,\hat{G}_K=G_K^0|\hat{K}=K\right)P\left(\hat{K}=K\right)\to 1\text{ as }T\to\infty.$$

(ii) Let $\mathcal{C}$ be any Borel-measurable set in $\mathbb{R}^p$. By (i),

$$\begin{aligned}P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k-\boldsymbol{\alpha}_k^0)\in\mathcal{C}\right) &= P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k-\boldsymbol{\alpha}_k^0)\in\mathcal{C}|\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{oracle}\right)P\left(\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{oracle}\right) \\ &\quad +P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k-\boldsymbol{\alpha}_k^0)\in\mathcal{C}|\hat{\boldsymbol{\beta}}\neq\hat{\boldsymbol{\beta}}^{oracle}\right)P\left(\hat{\boldsymbol{\beta}}\neq\hat{\boldsymbol{\beta}}^{oracle}\right) \\ &= P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0)\in\mathcal{C}\right)\{1-o(1)\}+o(1) \\ &\to P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0)\in\mathcal{C}\right)\text{ as }T\to\infty.\end{aligned}$$

That is, $\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k-\boldsymbol{\alpha}_k^0)$ shares the same asymptotic distribution as $\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0)$. As in the proof of Theorem 3.1, we have

$$\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0)=\left(\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_i\right)^{-1}\frac{1}{TN_k}\sum_{i\in G_k^0}\tilde{\mathbf{x}}_i'\boldsymbol{\varepsilon}_i.$$

By Assumption A4, (i) $\bar{\Phi}_k\equiv\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^T\tilde{\mathbf{x}}_{it}\tilde{\mathbf{x}}_{it}'\xrightarrow{P}\Phi_k>0$ and $\frac{1}{\sqrt{N_kT}}\sum_{i\in G_k^0}\sum_{t=1}^T\tilde{\mathbf{x}}_{it}\varepsilon_{it}-\mathbb{B}_{kNT}\xrightarrow{D}N(0,\Psi_k)$ as $(N_k,T)\to\infty$ or $T\to\infty$ alone. It follows that $\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle}-\boldsymbol{\alpha}_k^0)-\bar{\Phi}_k^{-1}\mathbb{B}_{kNT}\xrightarrow{D}N(0,\Phi_k^{-1}\Psi_k\Phi_k^{-1})$ and the conclusion in Theorem 3.2(ii) follows. ■

Proof of Theorem 3.3. Let $\mathcal{C}$ be defined as in the proof of Theorem 3.2(ii). In view of the fact

that $\hat{\boldsymbol{\alpha}}_{\hat{G}_k}$ becomes $\hat{\boldsymbol{\alpha}}_k^{oracle}$ conditional on $\hat{G}_k = G_k^0$, we have by Theorem 3.2(i)

$$\begin{aligned} P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0) \in \mathcal{C}\right) &= P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0) \in \mathcal{C} | \hat{G}_k = G_k^0\right) P\left(\hat{G}_k = G_k^0\right) \\ &\quad + P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0) \in \mathcal{C} | \hat{G}_k \neq G_k^0\right) P\left(\hat{G}_k \neq G_k^0\right) \\ &= P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle} - \boldsymbol{\alpha}_k^0) \in \mathcal{C}\right) \{1 - o(1)\} + o(1) \\ &\to P\left(\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle} - \boldsymbol{\alpha}_k^0) \in \mathcal{C}\right). \end{aligned}$$

That is, $\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0)$ is asymptotically equivalent to $\sqrt{N_kT}(\hat{\boldsymbol{\alpha}}_k^{oracle} - \boldsymbol{\alpha}_k^0)$ and the conclusion in Theorem 3.3 follows. ■

Proof of Theorem 3.4. The proof is built on and similar to that of Theorem 3.1 and we only sketch the main difference. The penalty term $P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ now becomes

$$P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \sum_{r=1}^{R} P_{\mathcal{B}_{\iota_r},\lambda_1,\lambda_2}(\boldsymbol{\beta}),$$

which involves the addition of R penalty terms. As assumed, $\{\mathcal{B}_{\iota_1}, ..., \mathcal{B}_{\iota_R}\}$ together forms an admissible segmentation net $\mathcal{N}$. For the first group G_1^0, there exists a $\mathcal{B}_{\iota_r} \in \mathcal{N}$ such that G_1^0 is properly segmented by $\mathcal{B}_{\iota_r}$. To make the notation easier to follow, we rename $\mathcal{B} = \mathcal{B}_{\iota_r}$ for the moment. Recall that $G_1^0 = \cup_{l=d_1}^{u_1} V_{1l}$, where $V_{1l} = G_1^0 \cap B_l$, and $B_l \in \mathcal{B}$. Without loss of generality and possibly with some renaming of notation, we can assume $B_{u_1} \backslash G_1^0 \neq \varnothing$ and $B_{u_1} \cap G_2^0 \neq \varnothing$. Here' $\backslash$' is the relative complement operator. Next we find the $\mathcal{B} \in \mathcal{N}$ that properly segments G_2^0. Similarly we can write $G_2^0 = \cup_{l=d_2}^{u_2} V_{2l}$. And so on. Finally, for each G_k^0 we have $G_k^0 = \cup_{l=d_k}^{u_k} V_{kl}$. The redefined segmentation $\mathcal{B}^* = \{V_{1d_1}, ..., V_{1u_1}, ..., V_{Kd_K}, ..., V_{Ku_K}\}$ is an admissible segmentation according to the definition. Now we decompose $P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ as

$$P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}) = P_{\mathcal{B}^*,\lambda_1,\lambda_2}(\boldsymbol{\beta}) + P_{\text{within}}(\boldsymbol{\beta}) + P_{\text{between}}(\boldsymbol{\beta}),$$

where $P_{\mathcal{B}^*,\lambda_1,\lambda_2}(\boldsymbol{\beta})$ is defined according to the new admissible segmentation $\mathcal{B}^*$, $P_{\text{within}}(\boldsymbol{\beta})$ contains all other penalty terms between members belonging to the same true group, and $P_{\text{between}}(\boldsymbol{\beta})$ contains all other penalty terms for members belonging to different true groups.

Next we specify the events.

1. Let $\mathcal{E}_1$ be the event that the segmentation net is admissible with the true parameters $\boldsymbol{\beta}^0$ so that we could generate the $\mathcal{B}^*$ described above. According to the assumption, we have $P(\mathcal{E}_1^c) \leq \epsilon_1$.

2. Let $\mathcal{E}_2 = \left\{\|\hat{\boldsymbol{\beta}}^{oracle} - \boldsymbol{\beta}^0\| \leq M \ln T \sqrt{K/T}\right\}$. According to the proof in Theorem 3.1, we have $P(\mathcal{E}_2^c) = o(K/T)$. Furthermore, over the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have property (i) in Theorem 3.1. Note that here $P_{\mathcal{B}^*,\lambda_1,\lambda_2}(\boldsymbol{\beta})$ plays a similar role to that of $P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ in Theorem 3.1; $P_{\text{within}}(\boldsymbol{\beta})$ and $P_{\text{between}}(\boldsymbol{\beta})$ are zero and a constant, respectively, conditional on $\mathcal{E}_1 \cap \mathcal{E}_2$.

3. Let $\mathcal{E}_3$ be as defined in Theorem 3.1 such that $P(\mathcal{E}_3^c) = o\left(T^{-1}\right)$. Combining the proof of Theorem 3.1 and arguments in the last point, we obtain a similar evaluation as property (ii) in Theorem 3.1.

Thus, just as in the proof of Theorem 3.1, we can show that, over the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, $\hat{\boldsymbol{\beta}}^{oracle}$ is the unique optimization solution of Q_{NT}. In addition, $P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - \epsilon_1 - o(K/T)$. ■

Supplementary Material for
"Homogeneity Pursuit in Panel Data Model: Theory and Applications"
(Not for publication in the main text of the paper)

Wuyi Wang[a], Peter C.B. Phillips[b], Liangjun Su[a]
[a] School of Economics, Singapore Management University
[b] Yale University, University of Auckland, University of Southampton, & Singapore Management University

This supplement states and proves a technical lemma that is used in the main text of the above paper.

B A Technical Lemma

Lemma B.1 *Let ξ_{it} denote a $d_\xi \times 1$ random vector with mean 0 and $\mathbb{E}\|\xi_{it}\|^q < \infty$ for some $q > 4$. Suppose that $\{\xi_{it}, i = 1, ..., N,\ t = 1, ..., T\}$ are independent across i and are strong mixing in the time index. Let $G_1^0, ..., G_K^0$ be defined as in the main text with $N_k = \left|G_k^0\right|$ for $k = 1, ..., K$. Let $\alpha_i(\cdot)$ denote the mixing coefficients of $\{\xi_{it}, t = 1, 2, ...\}$. Suppose that $\alpha_i(\tau) \leq \alpha(\tau)$ for all $i = 1, ..., N$ where $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $\rho \in (0, 1)$. Then as $T \to \infty$ and for some sufficiently large positive constant M and any positive constant c we have*

(i) $P\left(\left\|\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \xi_{it}\right\| \geq \frac{M\ln(N_k T)}{(N_k T)^{1/2}}\right) = o\left((N_k T)^{-1}\right)$ *for* $k = 1, ..., K$,

(ii) $P\left(\max_{1\leq i\leq N_k}\left\|\frac{1}{T}\sum_{t=1}^T \xi_{it}\right\| \geq \frac{M\ln(T)}{T^{1/2}}\right) = o\left(T^{-1}\right)$ *provided* $q > 8$ *and* $N_k = O\left(T^2\right)$,

(iii) $P\left(\max_{1\leq i\leq N}\left\|\frac{1}{T}\sum_{t=1}^T \xi_{it}\right\| \geq c\right) = o\left(T^{-1}\right)$ *provided* $N = O\left(T^2\right)$.

Proof. (i) Let $a_{N_k T} = M\ln(N_k T)/\sqrt{N_k T}$ and $\eta_{N_k T} = (N_k T)^\vartheta$ for $\vartheta = \frac{2}{q}$. Let ι_ξ be an arbitrary $d_\xi \times 1$ nonrandom vector with $\|\iota_\xi\| = 1$. Let $\mathbf{1}_{it} = \mathbf{1}\left\{\|\xi_{it}\| \leq \eta_{N_k T}\right\}$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$. Define

$$\xi_{1it} = \iota_\xi'\left[\xi_{it}\mathbf{1}_{it} - \mathbb{E}\left(\xi_{it}\mathbf{1}_{it}\right)\right],\ \xi_{2it} = \iota_\xi'\xi_{it}\bar{\mathbf{1}}_{it},\ \text{and}\ \xi_{3it} = \iota_\xi'\mathbb{E}\left(\xi_{it}\bar{\mathbf{1}}_{it}\right).$$

Apparently $\xi_{1it} + \xi_{2it} - \xi_{3it} = \iota_\xi'\xi_{it}$ as $\mathbb{E}(\xi_{it}) = 0$. We prove the lemma by showing that

$$\text{(i1) } N_k T \cdot P\left(\left|\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \xi_{1it}\right| \geq a_{N_k T}\right) = o(1),$$

$$\text{(i2) } N_k T \cdot P\left(\left|\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \xi_{2it}\right| \geq a_{N_k T}\right) = o(1),\ \text{and (i3)}\ \left|\frac{1}{N_k T}\sum_{i\in G_k^0}\sum_{t=1}^T \xi_{3it}\right| = o\left(a_{N_k T}\right).$$

First, we prove (i3). By the Hölder and Markov inequalities

$$\begin{aligned}\left|\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\xi_{3it}\right| &\leq \max_{1\leq i\leq N_k}\max_{1\leq t\leq T}\left\|\mathbb{E}\left(\xi_{it}\bar{\mathbf{1}}_{it}\right)\right\| \\ &\leq \max_{1\leq i\leq N_k}\max_{1\leq t\leq T}\left\{\mathbb{E}\left\|\xi_{it}\right\|^{q/2}\right\}^{2/q}\left\{P\left(\left\|\xi_{it}\right\|>\eta_{N_kT}\right)\right\}^{(q-2)/q} \\ &\leq c_{1q}\max_{1\leq i\leq N_k}\max_{1\leq t\leq T}\left\{P\left(\left\|\xi_{it}\right\|>\eta_{N_kT}\right)\right\}^{(q-2)/q} \\ &\leq c_{1q}\max_{1\leq i\leq N_k}\max_{1\leq t\leq T}\left\{\eta_{N_kT}^{-q}\mathbb{E}\left(\left\|\xi_{it}\right\|^{q}\right)\right\}^{(q-2)/q} \\ &= c_{1q}c_{2q}\eta_{N_kT}^{-(q-2)} = O\left(\left(N_kT\right)^{-\vartheta(q-2)}\right) = o(a_{N_kT}),\end{aligned}$$

where $c_{1q}\equiv\max_{i\in G_k^0}\max_{1\leq t\leq T}\left\{\mathbb{E}\left\|\xi_{it}\right\|^{q/2}\right\}^{2/q}$ and $c_{2q}\equiv\max_{i\in G_k^0}\max_{1\leq t\leq T}\left\{\mathbb{E}\left(\left\|\xi_{it}\right\|^{q}\right)\right\}^{(q-2)/q}$.

Next, we prove (i2). Noting that $\left\|\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\xi_{2it}\right\|\geq a_{N_kT}$ implies that $\max_{1\leq i\leq N_k}\max_{1\leq t\leq T}\left\|\xi_{it}\right\|>\eta_{N_kT}$, by the Boole and Markov inequalities, the dominated convergence theorem, and the stated conditions, we have

$$\begin{aligned}P\left(\left\|\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\xi_{2it}\right\|\geq a_{N_kT}\right) &\leq P\left[\max_{i\in G_k^0}\max_{1\leq t\leq T}\left\|\xi_{it}\right\|>\eta_{N_kT}\right] \\ &\leq \frac{N_kT}{\eta_{N_kT}^{q}}\max_{i\in G_k^0}\max_{1\leq t\leq T}\mathbb{E}\left[\left\|\xi_{it}\right\|^{q}\mathbf{1}\left\{\left\|\xi_{it}\right\|>\eta_{N_kT}\right\}\right] \\ &= o\left(\left(N_kT\right)^{1-q\vartheta}\right) = o\left(\left(N_kT\right)^{-1}\right).\end{aligned}$$

To prove (i1), we need to rewrite the expression $Q_{1NT}\equiv\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\xi_{1it}$. Without loss of generality, we assume that we can split the time interval $[1,T]$ into $2r_{N_kT}$ blocks with each block of length $l_{N_kT}=T/\left(2r_{N_kT}\right)\asymp\left(N_kT\right)^{\frac{1}{2}-\vartheta}$ where $a_T\asymp b_T$ means that a_T/b_T is bounded away from both 0 and infinity as $T\to\infty$. Then

$$\sum_{t=1}^{T}\xi_{1it}=\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}+\sum_{s=1}^{r_{N_kT}}B_{i,2s},$$

where $B_{i,s}=\frac{1}{N_kT}\sum_{t=(s-1)l_{N_kT}+1}^{sl_{N_kT}}\xi_{1it}$ for $s=1,...,2r_{N_kT}$. It follows that

$$P\left(\left|\frac{1}{N_kT}\sum_{i\in G_k^0}\sum_{t=1}^{T}\xi_{1it}\right|\geq a_{N_kT}\right)\leq P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}\right|\geq a_{N_kT}/2\right)+P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s}\right|\geq a_{N_kT}/2\right).$$

Below we show that the first term can be bounded by $o((N_kT)^{-1})$. The second term can be studied by using analogous arguments. Note that

$$\max_{i\in G_k^0}\max_{1\leq s\leq r_{N_kT}}\left|B_{i,2s-1}\right|=\frac{1}{N_kT}\max_{i\in G_k^0}\max_{1\leq s\leq r_{N_kT}}\left|\sum_{t=(2s-2)l_{N_kT}+1}^{(2s-1)l_{N_kT}}\iota_\xi'\left[\xi_{it}\mathbf{1}_{it}-\mathbb{E}\left(\xi_{it}\mathbf{1}_{it}\right)\right]\right|\leq\frac{2l_{N_kT}\eta_{N_kT}}{N_kT}\equiv C_{\xi N_kT}.$$

By the Davydov inequality, we can readily show that

$$\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}\mathbb{E}\left[\left(B_{i,2s-1}\right)^2\right]=\frac{1}{N_k^2T^2}\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}\mathbb{E}\left[\left(\sum_{t=(2s-2)l_{N_kT}+1}^{(2s-1)l_{N_kT}}\iota_\xi'\left[\xi_{it}\mathbf{1}_{it}-\mathbb{E}\left(\xi_{it}\mathbf{1}_{it}\right)\right]\right)^2\right]\leq\frac{C_1}{N_kT}$$

for some $C_1<\infty$. By Bradley's lemma (e.g., Lemma 1.2 in Bosq 1998), we can construct a sequence of random variables $B_{i,1}^*$, $B_{i,3}^*$,... such that (1) $B_{i,1}^*$, $B_{i,3}^*$,.. are independent, (2) $B_{i,2s-1}^*$ has the same distribution as $B_{i,2s-1}$, and (3) for any $C_2\in(0,C_{\xi N_kT}]$,

$$P\left\{\left|B_{i,2s-1}^*-B_{i,2s-1}\right|>C_2\right\}\leq 18(C_{\xi N_kT}/C_2)^{1/2}\alpha\left(l_{N_kT}\right). \tag{B.1}$$

Then we have

$$\begin{aligned}
& P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}\right|\geq a_{N_kT}/2\right)\\
\leq\ & P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}^*\right|\geq a_{N_kT}/4\right)+P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}\left(B_{i,2s-1}^*-B_{i,2s-1}\right)\right|\geq a_{N_kT}/4\right)\\
\equiv\ & I+II,\ \text{say.}
\end{aligned}$$

In view of the fact that $\exp(x)\leq 1+x+x^2$ for $|x|\leq 1/2$, $1+x\leq\exp(x)$ for any $x\geq 0$, and $\mathbb{E}\left[B_{i,2s-1}\right]=0$, we have for $\lambda_{N_kT}\equiv C_{\xi N_kT}^{-1}/2$,

$$\mathbb{E}\left[\exp\left(\pm\lambda_{N_kT}B_{i,2s-1}\right)\right]\leq 1+\lambda_{N_kT}^2\mathbb{E}\left[\left(B_{i,2s-1}\right)^2\right]\leq\exp\left(\lambda_{N_kT}^2\mathbb{E}\left[\left(B_{i,2s-1}\right)^2\right]\right).$$

Then by the Markov inequality, we have

$$\begin{aligned}
I\ =\ & P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}^*\right|\geq a_{N_kT}/4\right)\\
\leq\ & \exp\left(-\frac{\lambda_{N_kT}a_{N_kT}}{4}\right)\mathbb{E}\left\{\exp\left(\lambda_{N_kT}\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}^*\right)+\exp\left(-\lambda_{N_kT}\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}B_{i,2s-1}^*\right)\right\}\\
=\ & \exp\left(-\frac{\lambda_{N_kT}a_{N_kT}}{4}\right)\left\{\prod_{i\in G_k^0}\prod_{s=1}^{r_{N_kT}}\mathbb{E}\left[\exp\left(\lambda_{N_kT}B_{i,2s-1}^*\right)\right]+\prod_{i\in G_k^0}\prod_{s=1}^{r_{N_kT}}\mathbb{E}\left[\exp\left(-\lambda_{N_kT}B_{i,2s-1}^*\right)\right]\right\}\\
\leq\ & 2\exp\left(-\frac{\lambda_{N_kT}a_{N_kT}}{4}\right)\prod_{i\in G_k^0}\prod_{s=1}^{r_{N_kT}}\exp\left(\lambda_{N_kT}^2\mathbb{E}\left[\left(B_{i,2s-1}\right)^2\right]\right)\\
=\ & 2\exp\left(-\frac{\lambda_{N_kT}a_{N_kT}}{4}+\lambda_{N_kT}^2\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}\mathbb{E}\left[\left(B_{i,2s-1}\right)^2\right]\right)\\
\asymp\ & \exp\left(-M\ln\left(N_kT\right)\right)=o\left(\left(N_kT\right)^{-1}\right),
\end{aligned}$$

where the last line follows because $\lambda_{N_kT}^2/(N_kT)=\left(\frac{N_kT}{4l_{N_kT}\eta_{N_kT}}\right)^2/(N_kT)=\frac{N_kT}{16l_{N_kT}^2\eta_{N_kT}^2}\asymp l_{N_kT}^{-2}\left(N_kT\right)^{1-2\vartheta}\asymp 1$ and

$$\lambda_{N_kT}a_{N_kT}=\frac{N_kT}{4l_{N_kT}\eta_{N_kT}}\frac{M\ln\left(N_kT\right)}{\left(N_kT\right)^{1/2}}=\frac{M\left(N_kT\right)^{\frac{1}{2}-\vartheta}\ln\left(N_kT\right)}{4l_{N_kT}}\asymp M\ln\left(N_kT\right).$$

In addition, by (B.1) and the fact $\frac{a_{N_kT}}{4N_kr_{N_kT}} \leq C_{\xi N_kT}$

$$\begin{aligned} II &= P\left(\left|\sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}}\left(B^*_{i,2s-1}-B_{i,2s-1}\right)\right| \geq \frac{a_{N_kT}}{4}\right) \\ &\leq \sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}} P\left(\left|B^*_{i,2s-1}-B_{i,2s-1}\right| \geq \frac{a_{N_kT}}{4N_kr_{N_kT}}\right) \leq \sum_{i\in G_k^0}\sum_{s=1}^{r_{N_kT}} 18\left(\frac{C_{\xi N_kT}}{\frac{a_{N_kT}}{4N_kr_{N_kT}}}\right)^{1/2}\alpha\left(l_{N_kT}\right) \\ &= 36N_kr_{N_kT}\left(\frac{C_{\xi N_kT}N_kr_{N_kT}}{a_{N_kT}}\right)^{1/2}\alpha\left(l_{N_kT}\right) \leq \left(N_kT\right)^{-L} \text{ for sufficiently large } T, \end{aligned}$$

where L can be chosen arbitrarily large as $\alpha\left(l_{N_kT}\right)$ decays to zero at the exponential rate and $l_{N_kT} \asymp (N_kT)^{\frac{1-2\vartheta}{2}}$ diverges to ∞ at a polynomial rate.

This completes the proof of (i).

(ii) The proof is similar to that of (i) and is therefore sketched. Let $a_T = M\ln T/\sqrt{T}$ and $\eta_T = T^{\bar{\vartheta}}$ for $\bar{\vartheta} = \frac{4}{q}$. Let ι_ξ be an arbitrary $d_\xi \times 1$ nonrandom vector with $\|\iota_\xi\| = 1$. Let $\mathbf{1}_{it} = \mathbf{1}\{\|\xi_{it}\| \leq \eta_T\}$ and $\bar{\mathbf{1}}_{it} = 1 - \mathbf{1}_{it}$. Define $\bar{\xi}_{1it} = \iota'_\xi\left[\xi_{it}\mathbf{1}_{it} - \mathbb{E}\left(\xi_{it}\mathbf{1}_{it}\right)\right]$, $\bar{\xi}_{2it} = \iota'_\xi\xi_{it}\bar{\mathbf{1}}_{it}$, and $\bar{\xi}_{3it} = \iota'_\xi\mathbb{E}\left(\xi_{it}\bar{\mathbf{1}}_{it}\right)$. Apparently $\bar{\xi}_{1it} + \bar{\xi}_{2it} - \bar{\xi}_{3it} = \iota'_\xi\xi_{it}$ as $\mathbb{E}(\xi_{it}) = 0$. We prove the lemma by showing that

$$\begin{aligned} &\text{(ii1) } T\cdot P\left(\max_{i\in G_k^0}\left|\frac{1}{T}\sum_{t=1}^T\bar{\xi}_{1it}\right| \geq a_T\right) = o\left(1\right), \\ &\text{(ii2) } T\cdot P\left(\max_{i\in G_k^0}\left|\frac{1}{T}\sum_{t=1}^T\bar{\xi}_{2it}\right| \geq a_T\right) = o\left(1\right), \text{ and (ii3) } \max_{i\in G_k^0}\left|\frac{1}{T}\sum_{t=1}^T\bar{\xi}_{3it}\right| = o\left(a_T\right). \end{aligned}$$

Following the proof of (i3) and using the Hölder and Markov inequalities, we can readily show that

$$\max_{i\in G_k^0}\left|\frac{1}{T}\sum_{t=1}^T\bar{\xi}_{3it}\right| \leq \max_{i\in G_k^0}\max_{1\leq t\leq T}\left\|\mathbb{E}\left(\xi_{it}\bar{\mathbf{1}}_{it}\right)\right\| \leq c_{1q}c_{2q}\eta_T^{-(q-2)} = O\left(T^{-\bar{\vartheta}(q-2)}\right) = o(a_T).$$

Similarly, following the proof of (i2) and using the Boole and Markov inequalities, the dominated convergence theorem, and the stated conditions, we have

$$\begin{aligned} P\left(\max_{i\in G_k^0}\left|\frac{1}{T}\sum_{t=1}^T\bar{\xi}_{2it}\right| \geq a_{N_kT}\right) &\leq P\left[\max_{i\in G_k^0}\max_{1\leq t\leq T}\|\xi_{it}\| > \eta_{N_kT}\right] \\ &\leq \frac{N_kT}{\eta_T^q}\max_{i\in G_k^0}\max_{1\leq t\leq T}\mathbb{E}\left[\|\xi_{it}\|^q\mathbf{1}\left\{\|\xi_{it}\| > \eta_{N_kT}\right\}\right] \\ &= o\left(N_kT^{1-q\bar{\vartheta}}\right) = o\left(T^{-1}\right) \end{aligned}$$

where we use the fact that $N_k = O\left(T^2\right)$.

For (ii1), we assume that we can split the time interval $[1,T]$ into $2r_T$ blocks with each block of length $l_T = T/\left(2r_T\right) \asymp T^{\frac{1}{2}-\bar{\vartheta}}$. Then $\sum_{t=1}^T\bar{\xi}_{1it} = \sum_{s=1}^{r_T}\bar{B}_{i,2s-1} + \sum_{s=1}^{r_T}\bar{B}_{i,2s}$, where $\bar{B}_{i,s} =$

$\frac{1}{T}\sum_{t=(s-1)l_T+1}^{sl_T}\bar{\xi}_{1it}$ for $s=1,...,2r_T$. It follows that

$$P\left(\max_{i\in G_k^0}\left|\frac{1}{T}\sum_{t=1}^{T}\bar{\xi}_{1it}\right|\geq a_T\right)\leq P\left(\max_{i\in G_k^0}\left|\sum_{s=1}^{r_T}\bar{B}_{i,2s-1}\right|\geq a_T/2\right)+P\left(\max_{i\in G_k^0}\left|\sum_{s=1}^{r_T}\bar{B}_{i,2s}\right|\geq a_T/2\right).$$

Below we show that the first term can be bounded by $o\left(T^{-1}\right)$. The second term can be studied by using analogous arguments. Note that

$$\max_{i\in G_k^0}\max_{1\leq s\leq r_{N_kT}}\left|\bar{B}_{i,2s-1}\right|=\frac{1}{T}\max_{i\in G_k^0}\max_{1\leq s\leq r_T}\left|\sum_{t=(2s-1)l_T+1}^{2sl_T}\iota_\xi'\left[\xi_{it}\mathbf{1}_{it}-\mathbb{E}\left(\xi_{it}\mathbf{1}_{it}\right)\right]\right|\leq\frac{2l_T\eta_T}{T}\equiv\bar{C}_{\xi T}.$$

By the Davydov inequality, we can readily show that

$$\sum_{s=1}^{r_T}\mathbb{E}\left[\left(B_{i,2s-1}\right)^2\right]=\frac{1}{T^2}\sum_{s=1}^{r_T}\mathbb{E}\left[\left(\sum_{t=(2s-1)l_T+1}^{2sl_T}\iota_\xi'\left[\xi_{it}\mathbf{1}_{it}-\mathbb{E}\left(\xi_{it}\mathbf{1}_{it}\right)\right]\right)^2\right]\leq\frac{\bar{C}_1}{T}$$

for some $\bar{C}_1<\infty$. By Bradley's lemma, we can construct a sequence of random variables $\bar{B}_{i,1}^*$, $\bar{B}_{i,3}^*$, ... such that (1) $\bar{B}_{i,1}^*$, $\bar{B}_{i,3}^*$, .. are independent, (2) $\bar{B}_{i,2s-1}^*$ has the same distribution as $\bar{B}_{i,2s-1}$, and (3) for any $\bar{C}_2\in(0,\bar{C}_{\xi T}]$,

$$P\left\{\left|B_{i,2s-1}^*-B_{i,2s-1}\right|>\bar{C}_2\right\}\leq 18(\bar{C}_{\xi T}/\bar{C}_2)^{1/2}\alpha\left(l_T\right). \tag{B.2a}$$

Then we have

$$\begin{aligned}
& P\left(\max_{i\in G_k^0}\left|\sum_{s=1}^{r_T}B_{i,2s-1}\right|\geq a_T/2\right)\\
\leq\ & P\left(\max_{i\in G_k^0}\left|\sum_{s=1}^{r_T}B_{i,2s-1}^*\right|\geq a_T/4\right)+P\left(\max_{i\in G_k^0}\left|\sum_{s=1}^{r_T}\left(B_{i,2s-1}^*-B_{i,2s-1}\right)\right|\geq a_T/4\right)\\
\equiv\ & III+IV,\ \text{say.}
\end{aligned}$$

Noting that $\mathbb{E}\left[\exp\left(\pm\bar{\lambda}_T\bar{B}_{i,2s-1}\right)\right]\leq 1+\bar{\lambda}_T^2\mathbb{E}\left[\left(\bar{B}_{i,2s-1}\right)^2\right]\leq\exp\left(\bar{\lambda}_T^2\mathbb{E}\left[\left(\bar{B}_{i,2s-1}\right)^2\right]\right)$ for $\bar{\lambda}_T\equiv\bar{C}_{\xi T}^{-1}/2$ and by the Markov inequality, we have

$$\begin{aligned}
III\ &\leq\ \sum_{i\in G_k^0}P\left(\left|\sum_{s=1}^{r_T}\bar{B}_{i,2s-1}^*\right|\geq a_T/4\right)\leq 2\sum_{i\in G_k^0}\exp\left(-\frac{\bar{\lambda}_T a_T}{4}+\bar{\lambda}_T^2\sum_{s=1}^{r_T}\mathbb{E}\left[\left(\bar{B}_{i,2s-1}\right)^2\right]\right)\\
&\asymp\ \exp\left(-M\ln T\right)=o\left(T^{-1}\right)\ \text{for large } M,
\end{aligned}$$

where the last line follows because $\bar{\lambda}_T^2/T=\left(\frac{T}{4l_T\eta_T}\right)^2/T=\frac{T}{16l_T^2\eta_T^2}\asymp l_T^{-2}T^{1-2\vartheta}\asymp 1$ and $\bar{\lambda}_T a_T=\frac{T}{4l_T\eta_T}\frac{M\ln T}{T^{1/2}}=\frac{MT^{\frac{1}{2}-\vartheta}\ln T}{4l_T}\asymp M\ln T$.

In addition, by (B.2a) and the fact $\frac{a_T}{4r_T} \leq \bar{C}_{\xi T}$,

$$\begin{aligned} IV &= P\left(\max_{i\in G_k^0}\left|\sum_{s=1}^{r_T}\left(B^*_{i,2s-1}-B_{i,2s-1}\right)\right| \geq \frac{a_T}{4}\right) \\ &\leq \sum_{i\in G_k^0}\sum_{s=1}^{r_T} P\left(\left|B^*_{i,2s-1}-B_{i,2s-1}\right| \geq \frac{a_T}{4r_T}\right) \leq \sum_{i\in G_k^0}\sum_{s=1}^{r_T} 18\left(\frac{\bar{C}_{\xi T}}{\frac{a_T}{4r_T}}\right)^{1/2}\alpha\left(l_T\right) \\ &= 36N_k r_T\left(\frac{\bar{C}_{\xi T} r_T}{a_T}\right)^{1/2}\alpha\left(l_T\right) \leq T^{-L} \text{ for sufficiently large } T, \end{aligned}$$

where L can be chosen arbitrarily large. This completes the proof of (ii).

(iii) The proof is similar to (ii) and is again only sketched here. Let $a_T = c$ and $\eta_T = T^{\bar{\vartheta}}$ for $\bar{\vartheta} = \frac{4}{q}$. Let $\bar{\xi}_{1it}, \bar{\xi}_{2it}, \bar{\xi}_{3it}$, $\bar{B}_{i,s}$, $\bar{B}^*_{i,s}$, and $\bar{C}_{\xi T}$ be as defined in the proof of (ii). We prove the lemma by showing that

$$\begin{aligned} &\text{(iii1) } T\cdot P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\bar{\xi}_{1it}\right| \geq a_T\right) = o\left(1\right), \\ &\text{(iii2) } T\cdot P\left(\max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\bar{\xi}_{2it}\right| \geq a_T\right) = o\left(1\right), \text{ and (iii3) } \max_{1\leq i\leq N}\left|\frac{1}{T}\sum_{t=1}^{T}\bar{\xi}_{3it}\right| = o\left(1\right). \end{aligned}$$

The proofs of (iii2) and (iii3) are similar to those of (ii2) and (ii3) and omitted. For (iii1), we now assume that we can split the time interval $[1,T]$ into $2r_T$ blocks with each block of length $l_T = T/\left(2r_T\right) \asymp T^{1-\bar{\vartheta}-\epsilon}$ where ϵ is an arbitrarily small positive number such that $1-\bar{\vartheta}-\epsilon > 0$ (which is possible because $\bar{\vartheta} = \frac{4}{q} < 1$ under our assumption). Then $\sum_{t=1}^{T}\bar{\xi}_{1it} = \sum_{s=1}^{r_T}\bar{B}_{i,2s-1} + \sum_{s=1}^{r_T}\bar{B}_{i,2s}$,where $\bar{B}_{i,s} = \frac{1}{T}\sum_{t=(s-1)l_T+1}^{sl_T}\bar{\xi}_{1it}$ for $s = 1, ..., 2r_T$. Noting that $\mathbb{E}\left[\exp\left(\pm\bar{\lambda}_T\bar{B}_{i,2s-1}\right)\right] \leq 1+\bar{\lambda}_T^2\mathbb{E}\left[\left(\bar{B}_{i,2s-1}\right)^2\right] \leq \exp\left(\bar{\lambda}_T^2\mathbb{E}\left[\left(\bar{B}_{i,2s-1}\right)^2\right]\right)$ for $\bar{\lambda}_T \equiv \bar{C}_{\xi T}^{-1}/2$ and by the Markov inequality, we have

$$\begin{aligned} &P\left(\max_{1\leq i\leq N}\left|\sum_{s=1}^{r_T}B^*_{i,2s-1}\right| \geq a_T/4\right) \\ \leq\ & \sum_{i=1}^{N}P\left(\left|\sum_{s=1}^{r_T}\bar{B}^*_{i,2s-1}\right| \geq a_T/4\right) \leq 2\sum_{i=1}^{N}\exp\left(-\frac{\bar{\lambda}_T a_T}{4}+\bar{\lambda}_T^2\sum_{s=1}^{r_T}\mathbb{E}\left[\left(\bar{B}_{i,2s-1}\right)^2\right]\right) \\ \asymp\ & \exp\left(-cT^{\epsilon}+\ln N\right) = o\left(T^{-1}\right) \text{ for any } c>0 \text{ and } \epsilon > 0, \end{aligned}$$

where the last line follows because $\bar{\lambda}_T^2/T = \left(\frac{T}{4l_T\eta_T}\right)^2/T = \frac{T}{16l_T^2\eta_T^2} = O(l_T^2T^{1-2\vartheta}) = O\left(T^{-1+2\epsilon}\right) = o\left(1\right)$ for $\epsilon < 0.5$ and $\bar{\lambda}_T a_T = \frac{T}{4l_T\eta_T}c = cT^{\epsilon}$. In addition, as in the proof of (ii1), we can show that by Bradley's lemma, for sufficiently large T,

$$\begin{aligned} P\left(\max_{1\leq i\leq N}\left|\sum_{s=1}^{r_T}\left(B^*_{i,2s-1}-B_{i,2s-1}\right)\right| \geq a_T/4\right) &\leq \sum_{i=1}^{N}\sum_{s=1}^{r_T}18\left(\frac{\bar{C}_{\xi T}}{\frac{a_T}{4r_T}}\right)^{1/2}\alpha\left(l_T\right) \\ &= 36Nr_T\left(\frac{\bar{C}_{\xi T}r_T}{a_T}\right)^{1/2}\alpha\left(l_T\right) \leq T^{-L}, \end{aligned}$$

where L can be chosen arbitrarily large. The rest of the proof follows the corresponding part in the proof of (ii1). ■

References

Bosq, D. (1998), *Nonparametric statistics for stochastic processes: estimation and prediction.* Springer, New York.