# Homogeneity Pursuit in Panel Data Models: Theory and Application*

Wuyi Wang[a], Peter C.B. Phillips[b], Liangjun Su[c]
[a] Institute for Economic and Social Research, Jinan University
[b] Yale University, University of Auckland, University of Southampton, & Singapore Management University
[c] School of Economics, Singapore Management University

February 26, 2018

**Abstract**

This paper studies the estimation of a panel data model with latent structures where individuals can be classified into different groups with the slope parameters being 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. We show that it can identify the true group structure asymptotically and estimate the model parameters consistently at the same time. Simulations evaluate the performance and corroborate the asymptotic theory in several practical design settings. The empirical application reveals the heterogeneous grouping effect of income on democracy.

**JEL Classification:** C33, C38, C51

**Keywords:** CARDS; Clustering; Heterogeneous slopes; Income and democracy; Oracle estimator; Panel structure model

## 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.

*Correspondence should be addressed to Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903; Phone: +65 6828 0386. Email: peter.phillips@yale.edu (P.C.B. Phillips), ljsu@smu.edu.sg (L. Su), wuyi.wang.2013@phdecons.smu.edu.sg (W. Wang).

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 data models; see Lin and Ng (2012), Sarafidis and Weber (2015), Bonhomme and Manresa (2015), and Ando and Bai (2016). In addition, machine learning methods are also used to extract group patterns by using penalized extremum estimation. In particular, SSP (2016) develop classifier-Lasso (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 (2016) propose BIC-type information criteria to determine the number of groups. In addition, Lu and Su (2017) propose a direct testing procedure to identify the number of groups, and Su and Ju (2018) and Su, Wang, and Jin (2018) extend the C-Lasso method to nonparametric panels or panels with interactive fixed effects.

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 (see Bondell and Reich (2008) and Shen and Huang (2010)).

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 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 (2015) 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, KFW's (2015) 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 Penalized Least Squares 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 (2015) to the segmentation net, which enables us to extract groups more accurately. Just as CARDS for cross section data or the SSP (2016) C-Lasso for panel data, Panel-CARDS can identify the number of groups and estimate the parameters at the same time.

In comparison with existing methods in the literature, our methods have some distinctive characteristics. First, even though Lin and Ng (2012) and Sarafidis and Weber (2015) apply the K-means algorithm to study the panel structure model, they do not study the asymptotic properties of the resulting classification estimates. In contrast, we will study the asymptotic properties of the Panel-CARDS estimators. Bonhomme and Manresa (2015) and Ando and Bai (2016) adopt the K-means algorithm to study panel data models where the time or interactive fixed effects exhibit some group structure and study the asymptotic properties, but their models are different from the panel structure model considered here. Second, like SSP's (2016) C-Lasso method, our method is a Lasso-type penalization method, and the difference lies in the differences in the penalty term. Third, both K-means algorithm and C-Lasso methods require the specification of the number of groups while our Panel-CARDS method does not need to do so. In fact, existing theories for either

[1] An oracle estimate in this context is one that one can achieve the same asymptotic efficiency as if the exact group structure were known.

the K-means algorithm or C-Lasso method requires that the number of groups is fixed and the number of individuals within each group is proportional to the cross-sectional dimension, while our Panel-CARDS method allows the number of groups to pass to infinity at certain rate and the number of individuals within each group can be either divergent or fixed. In either case, we require $T \to \infty$, as often assumed in the literature. This largely broadens the scope of potential applications of our new method. Lastly, in comparison with the CARDS method, KFW (2015) require non-stochastic regressors and sub-Gaussian errors whereas we permit random regressors or lagged dependent variables, and replace the sub-Gaussian requirement by some moment conditions.

It is worth mentioning that like the early theoretical results in the literature, our asymptotic results are pointwise results. The implication is that in finite samples, the distributions of our estimators can be quite different from the normal, as discussed in Leeb and Pöscher (2005, 2008, 2009) and Schneider and Pöscher (2009). This is a well-known challenge in the literature of model selection no matter whether the selection is based on an information criterion or Lasso-type technique. Despite its importance, developing a thorough theory on uniform inference is beyond the scope of this paper.

We provide an empirical application of this new panel classification procedure. It 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 four latent groupings among the 74 countries corresponding to positive, negative, and indifferent associations between income and democracy.

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. Section 6 concludes. Proofs are given in the Appendix. The Online Supplement (Wang, Phillips, and Su 2018) provides additional technical material, proofs, convergence properties of the computational algorithm, some additional simulations, and further information on the empirical application.

*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 $A$. The symbol $\mathbf{1}\{\cdot\}$ denotes the indicator function. 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 the Panel-CARDS algorithm.

## 2.1 Panel structure models

Following SSP (2016), we consider a panel data model with latent group structure

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

where $\mathbf{x}_{it} = (x_{it1}, \ldots, x_{itp})'$ is a $p \times 1$ vector of regressors, $\mu_i$ is the individual fixed effect which may be correlated with $\mathbf{x}_{it}$, $\varepsilon_{it}$ 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

$$\boldsymbol{\beta}_i^0 = \sum_{k=1}^{K} \boldsymbol{\alpha}_k^0 \cdot \mathbf{1}\{i \in G_k^0\}. \tag{2.2}$$

Here $\boldsymbol{\alpha}_l^0 \neq \boldsymbol{\alpha}_k^0$ for any $l \neq k$, and $\mathcal{G} = \{G_1^0, G_2^0, \ldots, G_K^0\}$ forms a partition of $\{1, 2, \ldots, N\}$. Let $N_k = \left|G_k^0\right|$ denote the cardinality of $G_k^0$, $k = 1, \ldots, K$. Let $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_K')'$ and $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_N')'$. 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 Construction of the Panel-CARDS

This section describes how to construct the Panel-CARDS penalty function based on preliminary estimates of $\boldsymbol{\beta}_i^0$. Then we define a penalized least squares objective function.

### 2.2.1 Rank mapping in the panel data model

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

$$L_{NT}(\boldsymbol{\beta}) = \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i)^2, \tag{2.3}$$

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 $i = 1, 2, \ldots, N$.

Let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \tilde{\boldsymbol{\beta}}_2', \ldots, \tilde{\boldsymbol{\beta}}_N')'$ and $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \ldots, \tilde{\boldsymbol{\beta}}_N)$, which are $pN \times 1$ and $p \times N$ matrices, respectively. To use CARDS, we need to have a rank mapping over the cross section dimension according to the vector $\tilde{\boldsymbol{\beta}}$. If $p = 1$, the problem is exactly the same as the cross sectional case in KFW (2015). 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 ranking 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 ranking 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.

The three criteria connect the problem of ranking vectors with a famous impossibility theorem in social choice theory. In that setting, we take $\iota = 1, 2, \ldots, p$ 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 ranking of $N$ preliminary vector estimates (columns of matrix $\tilde{\mathbf{B}}$) must be totally determined by the ranking 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 $\iota^*$ from $\{1, 2, \ldots, p\}$ 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 $\iota^*$ denote the index of the regressor which has the largest variation across individuals for its coefficient estimates. Then we can sort $\{\tilde{\beta}_{i\iota^*}, i = 1, 2, \ldots, N\}$ to obtain the order

$$\tilde{\beta}_{\tau(1)\iota^*} \leq \tilde{\beta}_{\tau(2)\iota^*} \leq \cdots \leq \tilde{\beta}_{\tau(N)\iota^*}. \tag{2.4}$$

To proceed, we need to define an admissible segmentation**,** which is an ordered partition of a set.

**Definition 1.** *For a segmentation $\mathcal{B} = \{B_1, \ldots, B_L\}$ of the set $\{1, \ldots, N\}$ with true grouping structure $\mathcal{G} = \{G_1^0, G_2^0, \ldots, G_K^0\}$, 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 (KFW 2015) 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. It allows misclassification of individuals in the same group to different segments but only at the extent that they are still in "contiguous neighbor" sets.

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 (2.4) together with a tuning parameter $\delta$ 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.[2] But the segmentation $\mathcal{B} = \{B_1, \ldots, B_5\}$ 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 $\boldsymbol{\alpha}_1^0 = (1, 0.5)'$, $\boldsymbol{\alpha}_2^0 = (1, 1)'$, and $\boldsymbol{\alpha}_3^0 = (1, 1.5)'$. 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

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left( \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \right), \tag{2.5}$$

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 the next section.

### 2.2.2 Panel-CARDS objective function

Now suppose we have an admissible segmentation $\mathcal{B} = \{B_1, B_2, \ldots, B_L\}$. As in the KFW (2015) CARDS algorithm, we propose the following hybrid penalty

$$P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \underbrace{\sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}} p_{\lambda_1}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1)}_{\text{between-segment penalty}} + \underbrace{\sum_{l=1}^{L} \sum_{i \in B_l, j \in B_l} p_{\lambda_2}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1)}_{\text{within-segment penalty}}, \tag{2.6}$$

where $p_\lambda(\cdot)$ is the SCAD function of Fan and Li (2001).[3] The penalty function has two parts. The within-segment penalty drives slopes in the same segment to converge to each other when they are

[2] The value of $\delta$ determines the number of segments $L$ in $\mathcal{B}$. 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 $\tilde{\beta}_{1\iota^*} - \tilde{\beta}_{3\iota^*} > \delta$, $\cdots$ $\tilde{\beta}_{8\iota^*} - \tilde{\beta}_{10\iota^*} > \delta$ and $L = 5$. Besides, $V_{11} = \{2, 3\}$, $V_{12} = \{1\}$; $V_{22} = \{5\}$, $V_{23} = \{4, 6\}$; $V_{33} = \{7\}$, $V_{34} = \{9, 10\}$, $V_{35} = \{8\}$.

[3] The SCAD penalty function is given by

$$p_\lambda(x) = \begin{cases} \lambda|x| & \text{if } |x| \leq \lambda \\ -\frac{x^2 - 2a\lambda|x| + \lambda^2}{2(a-1)} & \text{if } \lambda < |x| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & \text{if } |x| > a\lambda \end{cases}$$

where $a > 2$. Following Fan and Li's (2001) recommendation, we set $a = 3.7$ in our simulations and applications.

actually in the same true group. The between-segment penalty 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. By adding the penalty term (2.6) to the original objective function (2.3), we obtain the following Penalized Least Squares objective function

$$Q_{NT}(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{2.7}$$

We call the above procedure basic Panel-CARDS. For implementation, we may apply the Local Linear Approximation algorithm to obtain the solution. We start from the initial solution and update it by solving the following iterative minimization problem

$$\hat{\boldsymbol{\beta}}^{(s+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ L_{NT}(\boldsymbol{\beta}) + R(\hat{\boldsymbol{\beta}}^{(s)}; \boldsymbol{\beta}) \right\}, \tag{2.8}$$

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 (2.8) is convex, we can apply a standard convex optimization package to obtain the solution. The justification of using Local Linear Approximation to solve (2.7) is relegated to Wang, Phillips, and Su (2018). 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 (2016), we can choose $\boldsymbol{\lambda} = (\delta, \lambda_1, \lambda_2)'$ to minimize the following information criterion

$$\mathrm{IC}(\boldsymbol{\lambda}) = \ln\left(\sigma_{NT}^2(\boldsymbol{\lambda})\right) + pK(\boldsymbol{\lambda})\rho_{NT}, \tag{2.9}$$

where $\sigma_{NT}^2(\boldsymbol{\lambda})$ and $K(\boldsymbol{\lambda})$ are estimates of $\sigma^2$ and the number of groups associated with $\boldsymbol{\lambda}$, and $\rho_{NT} = 0.5(NT)^{-1/2}$.[4] In Wang, Phillips, and Su (2018), we show this Information Criterion (IC) is effective in choosing tuning parameters.

This is a direct extension of CARDS from the cross sectional case to panel data. In this basic Panel-CARDS, 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.

In (2.5), no matter which regressor is used to construct the ordered segmentation, the original CARDS theory cannot work. Based on the first regressor, we are able to separate group 3 from the

[4] Too small or too large a $\delta$ 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 $\lambda_1$ and $\lambda_2$ depend on the value of coefficients we use in the DGP. Generally speaking, when the coefficients are large, the tuning parameters $\lambda_1$ and $\lambda_2$ are large correspondingly.

other two groups; and based on the second (or third) regressor, we can separate group 2 (or 3) from the other groups. This motivates us to propose the following concept of admissible segmentation net.

**Definition 2.** *Let $\mathcal{G} = \{G_1^0, G_2^0, \ldots, G_K^0\}$ denote the true grouping structure. Given $R$ segmentations $\mathcal{B}_{\iota_1}, \ldots, \mathcal{B}_{\iota_R}$, 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 $\mathcal{N} \equiv \{\mathcal{B}_{\iota_1}, \ldots, \mathcal{B}_{\iota_R}\}$ is called an admissible segmentation net.*

Note that the admissible segmentation does not always exist. In such cases, the admissible segmentation net, as a collection of different partitions of different ordered sets, plays an important role. 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 above basic Panel-CARDS procedure. Given an admissible segmentation net $\mathcal{N} = \{\mathcal{B}_{\iota_1}, \ldots, \mathcal{B}_{\iota_R}\}$, 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 (2.6).
- 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 Penalized Least Squares function:

$$Q_{NT}^*(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{2.10}$$

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 2015); and not too few, in order to handle the sparse parameters case introduced at the end of Section 2.2.1.[5]

There are two possible ways to choose $R < p$ regressors, based on how the segmentations are generated: (i) From the preliminary estimates we calculate the empirical variance of the slope coefficient estimates for each regressor $j$ (from 1 to $p$). That is, calculate the sample variance of $\{\tilde{\beta}_{1j}, \ldots, \tilde{\beta}_{nj}\}$ for $j$ from 1 to $p$. Then choose the $R$ regressors with the largest $R$ cross-sectional heterogeneity in the slope estimates. (ii) In applications, we may choose the $R$ regressors which are most likely to have heterogeneous responses. For further explanations, see Section E of Wang, Phillips, and Su (2018).

Although in the definition 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 reasonably well in finite samples.

[5] Its existence follows directly from Theorem 3 of KFW (2015).

### 2.3 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 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 of 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 hierarchical clustering is used here to improve the finite sample performance and its asymptotic theory can be justified provided such a prior information is correctly specified. The details of hierarchical clustering 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 $\tilde{\mathbf{x}}_i = (\tilde{\mathbf{x}}_{i1}, \ldots, \tilde{\mathbf{x}}_{iT})'$, $\tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, \ldots, \tilde{y}_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})'$, and $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$. 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 $\bar{\rho}_j(s) = \rho_j'(s) = p_{\lambda_j}'(|s|)\mathrm{sgn}(s)$ 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 $\{B_1, \ldots, B_L\}$, we define $\phi_k = N_k / \min\{N_k^3, \min_{d_k\leq l\leq u_k} |B_l|^2\}$. Note that $1/N_k^2 \leq \phi_k \leq N_k$. Let $\hat{\mathcal{G}}_{\hat{K}} = \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{\hat{K}}\}$ be an arbitrary partition of $\{1, \ldots, N\}$ where $|\hat{G}_k| \geq 1$ for $k = 1, \ldots, \hat{K}$. Define $\hat{\sigma}^2_{\hat{\mathcal{G}}_{\hat{K}}} = (NT)^{-1} \sum_{k=1}^{\hat{K}} \sum_{i\in\hat{G}_k} \sum_{t=1}^{T} (\tilde{y}_{it} - \tilde{x}_{it}'\check{\beta}_i)^2$, where $\{\check{\beta}_i\}$ solves the minimization problem with objective function $L_{NT}(\beta)$ and the constraint imposed by the group structure $\hat{\mathcal{G}}_{\hat{K}}$. 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 $i$, $\{(\mathbf{x}_{it}, y_{it}) : t = 1, 2, \ldots\}$ 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* $k = 1, \ldots, K$, $N_k$ *either passes to infinity or stays fixed as* $T \to \infty$, *and* $N = O\left(T^2\right)$.

**Assumption A2.** $p_\lambda(\cdot)$ *is a 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* $\lambda_1$ *and* $\lambda_2$ *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 $N_{\min} = \min\{N_1, \ldots, N_K\}$.

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

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

**Assumption A5.** (i) As $(N, T) \to \infty$**,** $\min_{1\leq \hat{K}<K}\min_{\hat{G}_{\hat{K}}}\hat{\sigma}^2_{\hat{\mathcal{G}}_{\hat{K}}} \xrightarrow{P} \bar{\sigma}^2 > \sigma_0^2$, where $\sigma_0^2 \equiv \lim_{(N,T)\to\infty}(NT)^{-1}\sum_{k=1}^{K^0}\sum_{i\in G_k^0}\sum_{t=1}^T \mathbb{E}(\tilde{y}_{it} - \tilde{x}'_{it}\beta_i^0)^2$.

(ii) As $(N, T) \to \infty$**,** $\rho_{NT} \to 0$ and $NT\rho_{NT} \to \infty$.

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 $\{\varepsilon_{it}, t = 1, 2, \ldots\}$ or $\{\mathbf{x}_{it}\varepsilon_{it}, t = 1, 2, \ldots\}$. 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 $\varepsilon_{it}$. In comparison with conditions 1 and 3 in KFW (2015) 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. It is possible that $N_k$'s are all fixed as $T \to \infty$. In contrast, SSP (2016) 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 (2015). Following KFW (2015), we specify $p_\lambda(\cdot)$ as the SCAD penalty function in our simulations and the application below. Assumption A3 imposes conditions on $K$, $b_{NT}$, $\lambda_1$ and $\lambda_2$. 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. It is much weaker than the separation requirement in Bonhomme and Manresa (2015) and SSP (2016). A3(ii) specifies the ranges of speed at which $\lambda_1$ and $\lambda_2$ shrink to zero. Assumption A4 is borrowed from SSP (2016) 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 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 (2016) for detailed discussions concerning A4. Assumption A5 is imposed to ensure the asymptotic validity of our information criterion (2.9). Assumption A5(i) assumes that for all under-fitted models, the mean square errors would be asymptotically greater than $\sigma_0^2$, and Assumption A5(i) is imposed to avoid both over- and under-fitted models.

### 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 $\mathcal{G} = \{G_1^0, \ldots, G_K^0\}$ 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 (2.3). 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

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

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 $\delta$ 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 (2.7) has a strictly local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \ldots, \hat{\boldsymbol{\beta}}_N')'$ 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 (2015). 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 $P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}) \to 1$ 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 $\hat{G}_k$, $k = 1, 2, \ldots, \hat{K}$ to denote the $\hat{K}$ estimated groups.

Let $\hat{\boldsymbol{\alpha}}_k$, $k = 1, 2, \ldots, \hat{K}$, denote the group-specific estimated slope coefficients. By definition,

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

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

[6] See Theorem 3 in KFW (2015) for a proof.

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

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

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 (2016) 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. Despite this fact, the success of Panel-CARDS hinges on the accuracy of preliminary estimates. Although Panel-CARDS is robust to mildly misranking of the preliminary estimates, poor preliminary estimates would deteriorate the performance of Panel-CARDS. Accordingly, one should be cautious about factors that affect the accuracy of preliminary estimates such as small $T$, low signal to noise ratio and too many regressors.

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

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

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* $k = 1, \ldots, K$, $\sqrt{N_k T}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0) - \bar{\Phi}_k^{-1}\mathbb{B}_{kNT} \xrightarrow{D} N(0, \Phi_k^{-1}\Psi_k\Phi_k^{-1})$ *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. Belloni and Chernozhukov (2013) show that the post-Lasso estimator performs at least as well as a Lasso estimator in terms of rate of convergence, and it has a smaller second-order bias 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 (2016) 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 the 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 $\delta$, 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 (2.10) 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.

Because of the limitations of the basic version of Panel-CARDS, we will use Panel-CARDS to denote the advanced version in the simulations and application below unless otherwise stated.

### 3.4 Analysis of Panel-CARDS with both individual and time fixed effects

In this subsection we consider the panel structure model with both individual and time fixed effects,

$$y_{it}=\mathbf{x}_{it}'\boldsymbol{\beta}_i+\mu_i+\gamma_t+\varepsilon_{it},\ i=1,\ldots,N,\ t=1,\ldots,T,$$

where $\gamma_t$ is the time fixed effect, all other variables are defined as above, and $\boldsymbol{\beta}_i$s have the latent group structure defined in (2.2). We study the asymptotic properties of Panel-CARDS estimators under this model.

As before, we first concentrate out the individual fixed effects to obtain

$$\tilde{y}_{it}=\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i+\tilde{\gamma}_t+\tilde{\varepsilon}_{it},$$

where $\tilde{\gamma}_t=\gamma_t-T^{-1}\sum_{s=1}^T\gamma_s$ and $\tilde{\varepsilon}_{it}=\varepsilon_{it}-T^{-1}\sum_{s=1}^T\varepsilon_{is}$. Then we get rid of $\tilde{\gamma}_t$ from the above equation to obtain

$$\ddot{y}_{it}=\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i-\frac{1}{N}\sum_{j=1}^N\tilde{\mathbf{x}}_{jt}'\boldsymbol{\beta}_j+\ddot{\varepsilon}_{it},$$

where $\ddot{y}_{it}=\tilde{y}_{it}-N^{-1}\sum_{j=1}^N\tilde{y}_{jt}$ and $\ddot{\varepsilon}_{it}=\tilde{\varepsilon}_{it}-N^{-1}\sum_{j=1}^N\tilde{\varepsilon}_{jt}$. Without knowing the latent group structure, we have the following objective function:

$$L_{2,NT}(\boldsymbol{\beta})=\frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T\left(\ddot{y}_{it}-\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i+\frac{1}{N}\sum_{j=1}^N\tilde{\mathbf{x}}_{jt}'\boldsymbol{\beta}_j\right)^2. \tag{3.4}$$

By minimizing the above objective function, we get the preliminary estimator $\tilde{\boldsymbol{\beta}}= (\tilde{\boldsymbol{\beta}}_1', \tilde{\boldsymbol{\beta}}_2', ..., \tilde{\boldsymbol{\beta}}_N')'$. The penalized least squares objective function is constructed as

$$Q^*_{2,NT}(\boldsymbol{\beta}) = L_{2,NT}(\boldsymbol{\beta}) + P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}), \tag{3.5}$$

where $P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ is as defined in (2.10). By solving (3.5) we obtain the Panel-CARDS estimator $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \ldots, \hat{\boldsymbol{\beta}}_N')'$. Let $\hat{\boldsymbol{\beta}}^{oracle}$ denote the oracle estimator of $\boldsymbol{\beta}$ by knowing the true group structure of $\boldsymbol{\beta}_i$s in (2.2). Let $\hat{\boldsymbol{\alpha}}^{oracle} = (\hat{\boldsymbol{\alpha}}_1^{oracle\prime}, ..., \hat{\boldsymbol{\alpha}}_K^{oracle\prime})'$ denote the group-specific version of $\hat{\boldsymbol{\beta}}^{oracle}$.

The following theorem reports the asymptotic properties of the Panel-CARDS estimator $\hat{\boldsymbol{\beta}}$ when both individual and time fixed effects appear.

**Theorem 3.5** *Suppose that Assumptions A1-A3 hold. Suppose that the preliminary estimate $\tilde{\boldsymbol{\beta}}$, the tuning parameter $\delta$, 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 in (3.5) has a strictly local minimizer $\hat{\boldsymbol{\beta}}$ such that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}$ and $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\| = O_p(\sqrt{K/T})$.*

This theorem shows that when the time fixed effects are added to our model, the Panel-CARDS still gives the oracle estimator with high probability. For inference, one needs to know the asymptotic distribution of $\hat{\boldsymbol{\alpha}}^{oracle}$; see Theorem 4.3 in Lu and Su (2017).

# 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 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 $\mu_i$ 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 $\boldsymbol{\alpha}_1^0 = (1, 2)'$, $\boldsymbol{\alpha}_2^0 = (1, 1)'$, and $\boldsymbol{\alpha}_3^0 = (2, 1)'$. 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.

**DGP 2**. Here we use DGP 1 in SSP (2016). 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 $\boldsymbol{\alpha}_1^0 = (0.4, 1.6)'$, $\boldsymbol{\alpha}_2^0 = (1, 1)'$, and $\boldsymbol{\alpha}_3^0 = (1.6, 0.4)'$.

**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 $\boldsymbol{\alpha}_1^0 = (0.6, 1.5, -1)'$, $\boldsymbol{\alpha}_2^0 = (0.6, 1, 0)'$, and $\boldsymbol{\alpha}_3^0 = (0.6, 0.5, 1)'$.

In DGPs 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 200 for all scenarios in this investigation.

### 4.2 Implementation and evaluation

Since the performance of the basic Panel-CARDS is not robust, we only implement the advanced Panel-CARDS in simulations. Recall that $\eta$ 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 $\eta$, 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 $\mathcal{A}^0 = \{A_1, A_2, \ldots, A_{\hat{K}^0}\}$, 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 $\mathcal{A} = \{A_1, A_2, \ldots, A_{\hat{K}}\}$, 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

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

Now we re-classify the element $j$ to group $A_{k^*}$ for $k^* \in \{1, \ldots, \hat{K}\}$. 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 $\widehat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{\hat{K}}\}$.

We use a BIC-type information criteria to choose the tuning parameters. Given the Panel-CARDS classification results $\widehat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, \ldots, \hat{G}_{\hat{K}}\}$, which are obtained by using the tuning parameter vector $\boldsymbol{\lambda}$, we calculate $\text{IC}(\boldsymbol{\lambda}) = \ln\left(\sigma^2_{NT}(\boldsymbol{\lambda})\right) + p\hat{K}/(2\sqrt{NT})$, where $\sigma^2_{NT}(\boldsymbol{\lambda}) = \frac{1}{NT}\sum_{s=1}^{\hat{K}} \sum_{i\in A_s}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}'_{it}\hat{\boldsymbol{\beta}}_s(\boldsymbol{\lambda}))^2$, 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 200 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 (2015), we use the Normalized Mutual Information 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 $\mathcal{A} = \{A_1, A_2, \ldots\}$ and $\mathcal{B} = \{B_1, B_2, \ldots\}$ on the same set $\{1, 2, \ldots, N\}$, the Normalized Mutual Information is defined as $\text{NMI}(\mathcal{A}, \mathcal{B}) = I(\mathcal{A}, \mathcal{B})/\sqrt{H(\mathcal{A})H(\mathcal{B})}$, 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 Normalized Mutual Information value is to 1. We report $\text{NMI}(\widehat{\mathcal{G}}, \mathcal{G})$ for all DGPs.

In addition, we report the correct classification ratio, root mean square error (RMSE), average bias (Bias), and coverage probability of the two-sided nominal 95% confidence intervals when $\eta = 2\%$. We follow SSP (2016) to define these criteria. The correct classification ratio is defined as $N^{-1}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k} 1\{\beta_i^0 = \alpha^0_{m(i)}\}$, where $m(i)$ denotes $i$'s true group member. For the last 3 criteria, we focus on the estimates of the second slope coefficients. Let $\alpha^0_{\cdot 2} \equiv (\alpha^0_{1,2}, \ldots, \alpha^0_{K^0,2})'$ denote the vector of the second regressor's slope coefficient of all groups. The RMSE is defined as the weighted average RMSEs of estimates of $\alpha^0_{k,2}$s with weights $N_k/N$: $\sum_{k=1}^{K}\frac{N_k}{N}\text{RMSE}(\alpha^0_{k,2})$. Similarly, we can define the average bias and the coverage probability for the 95% confidence intervals.

### 4.3 Simulation results

We will focus on the performance of the advanced Panel-CARDS. We use $R = 2$ regressors to construct the segmentation net. Given the matrix of preliminary estimates, $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \ldots, \tilde{\boldsymbol{\beta}}_N)$, 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.

Figure 1 reports the classification results for DGP 1 for different combinations of $N$, $T$, and $\eta$. It shows the Normalized Mutual Information between the estimated group structure $\widehat{\mathcal{G}}$ and the true group structure $\mathcal{G}$ and suggests that as $T$ increases, the Normalized Mutual Information 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 $\eta$. We also note that the performance of Panel-CARDS with $\eta = 2\%$ or $5\%$ significantly improves that with $\eta = 0$, but a further increase of $\eta$ does not necessarily lead to improved performance. Figure 2 reports the Normalized Mutual Information for DGP 2 for various combinations of $N$, $T$, and $\eta$. The Normalized Mutual Information patterns in Figure 2 are similar to those in Figure 1 for DGP 1. With respect to $\eta$, we also find that a choice of $\eta = 2\%$ or $5\%$ tends to outperform $\eta = 0$. Figure 3 shows the classification results for DGP 3 where the true number of groups is reasonably large (8 here). It demonstrates 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. Figure 4 reports the classification results for DGP 4 where the panel is a dynamic one. Apparently, the Panel-CARDS performs very well in this situation unless $T$ is very small and $\eta = 0$. The general conclusions from DGPs 1–3 also hold here. For the frequency of obtaining the estimated number of groups for all DGPs, see Section E in Wang, Phillips, and Su (2018).

For the second slope coefficients $\{\alpha_{k,2}^0\}_{k=1}^K$ and $\eta = 2\%$, Table 1 reports the correct classification ratio, RMSE, Bias, and 95% coverage probability of Panel-CARDS in Columns 4–7, and the RMSE, Bias, and 95% coverage probability of the oracle ones in Columns 8–10. For DGP 4, the estimators are bias-corrected by using the half-panel jackknife method of Dhaene and Jochmans (2015). As expected, the Panel-CARDS may not perform well when $T$ is small (10 or 20) in terms of correct classification ratio or coverage probability. But the performance of Panel-CARDS improves quickly as $T$ increases and appears almost as good as the oracle estimate when $T = 40$ or 80.

# 5 Empirical Application: Income and Democracy

## 5.1 Model and data

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 be similar or quite different. In South Korea, the degree of democracy increases when the economy is growing steadily. Similar patterns emerge for other countries such as Spain and Romania. However, for China the story is quite different. The democracy index composed by the Freedom House has not changed very much over the last four decades or so for China despite the fact that China has made remarkable economic progress over the same period. Moreover, for some countries like South Africa and Malaysia, a negative correlation is observed between income and democracy.

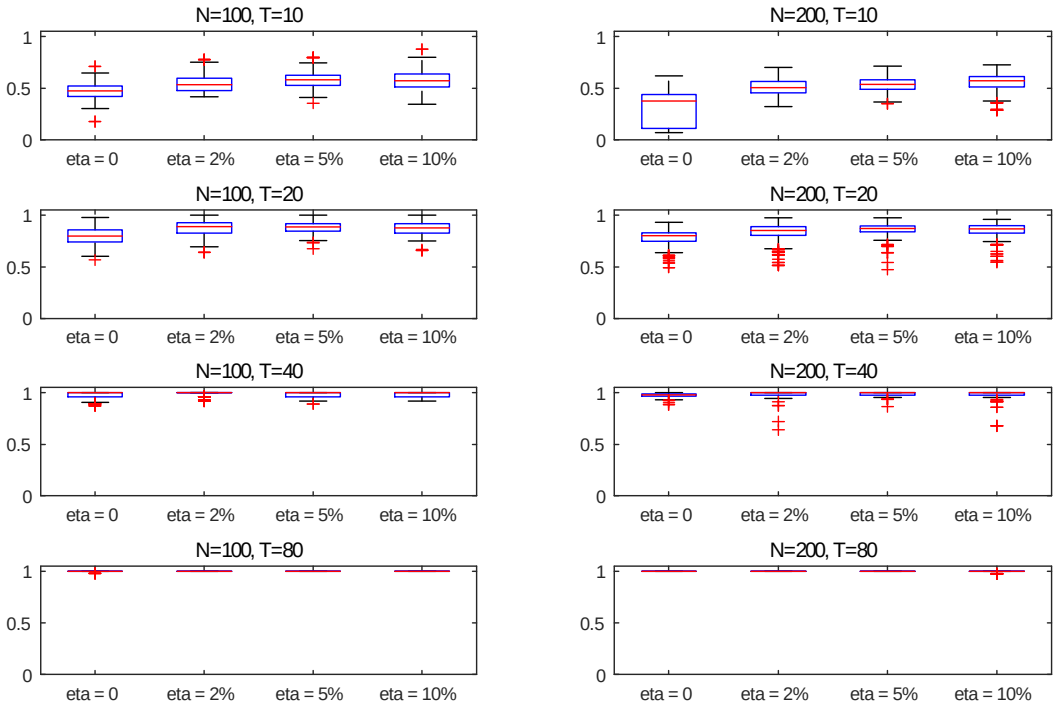


Figure 1: NMI of DGP 1 classification results using Panel-CARDS. The $x$-axis and $y$-axis mark the $\eta$ and NMI values, respectively. The left and right columns report the results for $N = 100$ and $N = 200$, respectively.

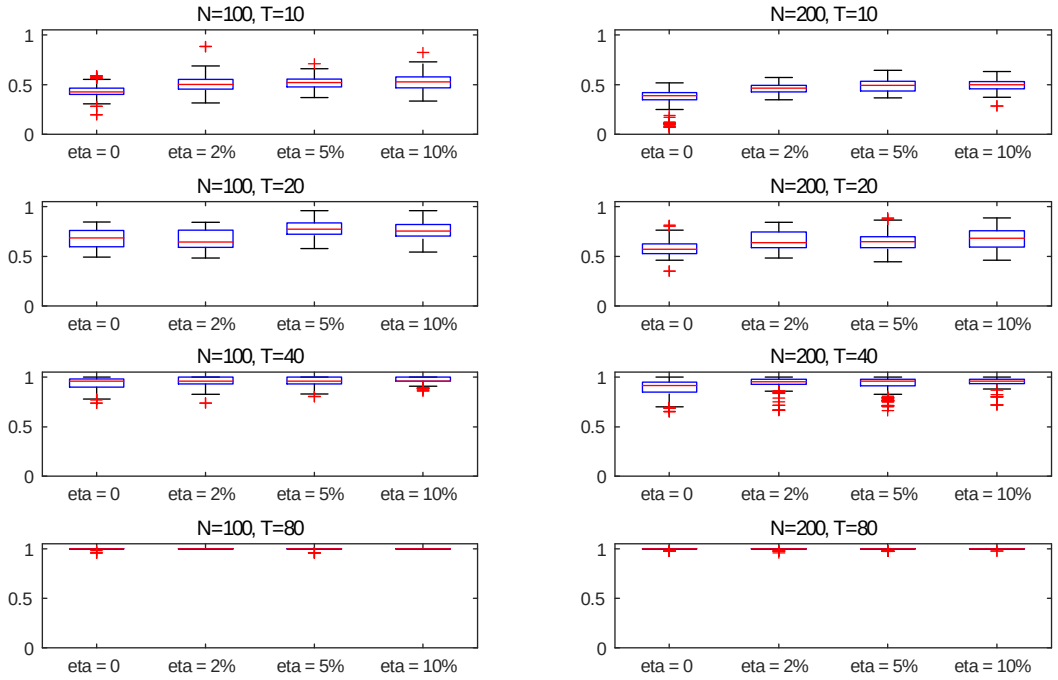


Figure 2: NMI of DGP 2 classification results using Panel-CARDS. (See Figure 1 for explanations.)

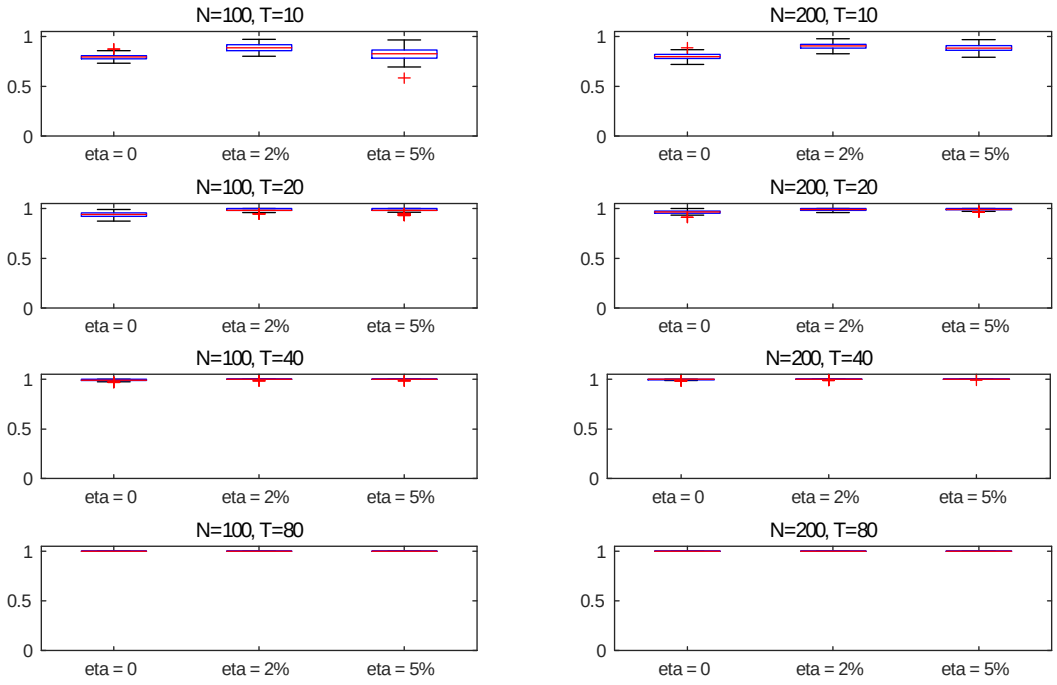


Figure 3: NMI of DGP 3 classification results using Panel-CARDS. (See Figure 1 for explanations.)

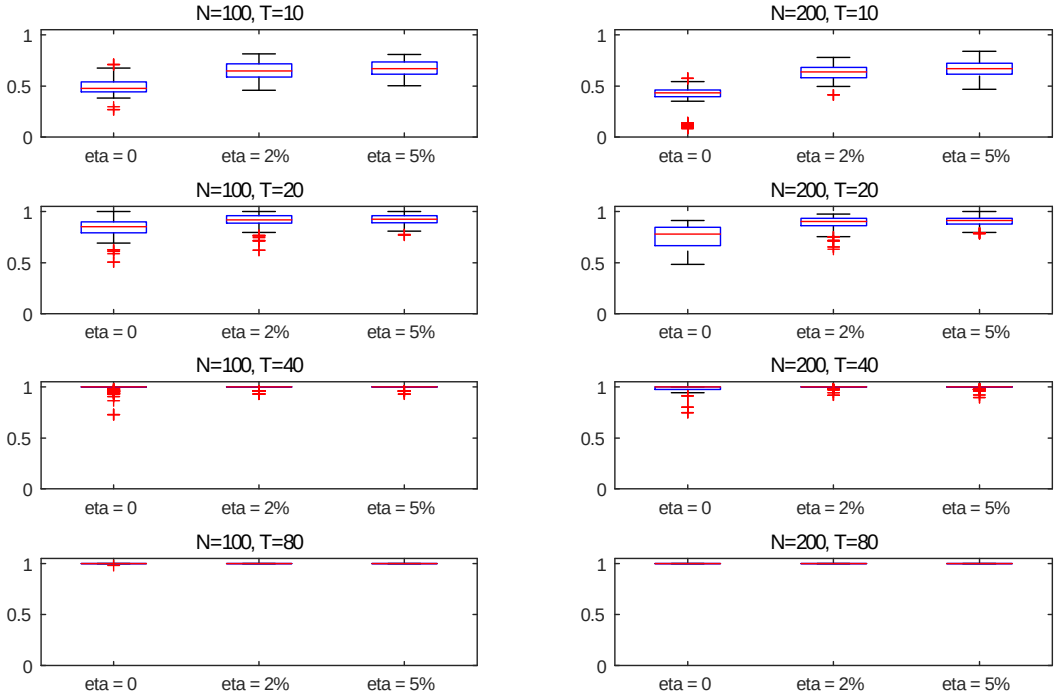


Figure 4: NMI of DGP 4 classification results using Panel-CARDS. (See Figure 1 for explanations.)

Table 1: Correct classification of individuals and point estimation of $\boldsymbol{\alpha}_{\cdot 2}^0$.

| | | | Panel-CARDS | | | | Oracle | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | $N$ | $T$ | % of Correct Classification | RMSE | Bias | Coverage | RMSE | Bias | Coverage |
| 1 | 100 | 10 | 0.714 | 0.425 | 0.016 | 0.582 | 0.080 | -0.002 | 0.941 |
| | 100 | 20 | 0.901 | 0.239 | -0.009 | 0.758 | 0.053 | 0.002 | 0.937 |
| | 100 | 40 | 0.988 | 0.091 | 0.003 | 0.934 | 0.038 | 0.003 | 0.956 |
| | 100 | 80 | 1 | 0.027 | -0.001 | 0.957 | 0.027 | -0.001 | 0.957 |
| | 200 | 10 | 0.683 | 0.426 | -0.014 | 0.412 | 0.056 | 0.005 | 0.910 |
| | 200 | 20 | 0.807 | 0.286 | -0.054 | 0.557 | 0.040 | 0.002 | 0.936 |
| | 200 | 40 | 0.963 | 0.088 | 0.005 | 0.906 | 0.027 | -0.001 | 0.947 |
| | 200 | 80 | 1.000 | 0.020 | 0.000 | 0.946 | 0.018 | -0.000 | 0.946 |
| 2 | 100 | 10 | 0.738 | 0.440 | -0.012 | 0.644 | 0.078 | -0.003 | 0.948 |
| | 100 | 20 | 0.956 | 0.195 | -0.005 | 0.903 | 0.054 | -0.002 | 0.955 |
| | 100 | 40 | 0.997 | 0.053 | 0.000 | 0.962 | 0.036 | 0.000 | 0.965 |
| | 100 | 80 | 1 | 0.025 | -0.000 | 0.965 | 0.025 | -0.000 | 0.965 |
| | 200 | 10 | 0.712 | 0.444 | -0.023 | 0.526 | 0.058 | 0.001 | 0.908 |
| | 200 | 20 | 0.939 | 0.226 | -0.041 | 0.841 | 0.039 | -0.002 | 0.942 |
| | 200 | 40 | 0.991 | 0.058 | -0.006 | 0.939 | 0.025 | -0.001 | 0.950 |
| | 200 | 80 | 1 | 0.019 | -0.001 | 0.954 | 0.019 | -0.001 | 0.954 |
| 3 | 100 | 10 | 0.886 | 0.395 | 0.357 | 0.520 | 0.137 | 0.002 | 0.931 |
| | 100 | 20 | 0.987 | 0.137 | 0.072 | 0.887 | 0.089 | 0.002 | 0.953 |
| | 100 | 40 | 1.000 | 0.067 | 0.001 | 0.938 | 0.065 | 0.001 | 0.938 |
| | 100 | 80 | 1 | 0.044 | -0.000 | 0.956 | 0.044 | -0.000 | 0.956 |
| | 200 | 10 | 0.923 | 0.335 | 0.256 | 0.685 | 0.093 | -0.005 | 0.943 |
| | 200 | 20 | 0.995 | 0.110 | -0.000 | 0.951 | 0.063 | -0.001 | 0.952 |
| | 200 | 40 | 1.000 | 0.048 | -0.002 | 0.932 | 0.046 | -0.002 | 0.932 |
| | 200 | 80 | 1 | 0.032 | 0.001 | 0.950 | 0.032 | 0.001 | 0.950 |
| 4 | 100 | 10 | 0.809 | 0.417 | -0.017 | 0.618 | 0.117 | -0.014 | 0.936 |
| | 100 | 20 | 0.966 | 0.177 | -0.005 | 0.916 | 0.065 | -0.003 | 0.953 |
| | 100 | 40 | 0.999 | 0.053 | -0.005 | 0.933 | 0.044 | -0.006 | 0.936 |
| | 100 | 80 | 1 | 0.030 | -0.002 | 0.956 | 0.030 | -0.002 | 0.956 |
| | 200 | 10 | 0.819 | 0.407 | -0.025 | 0.669 | 0.096 | -0.013 | 0.939 |
| | 200 | 20 | 0.949 | 0.186 | -0.004 | 0.883 | 0.056 | -0.005 | 0.948 |
| | 200 | 40 | 0.999 | 0.043 | -0.002 | 0.950 | 0.034 | -0.002 | 0.957 |
| | 200 | 80 | 1 | 0.021 | -0.007 | 0.965 | 0.021 | -0.007 | 0.965 |

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, BM hereafter), we consider the following regression model with both individual and time fixed effects:

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

where $d_{it}$ denotes a measure of democracy for country $i$ in period $t$ that is normalized to take values between 0 and 1, $I_{it}$ denotes the logarithm of the real GDP per capita for country $i$ in period $t$, $\mu_i$ is the individual fixed effect, $\gamma_t$ is the time fixed effect, $\varepsilon_{it}$ is the error term, and $\beta_{i1}$ and $\beta_{i2}$ are the slope coefficients, which are assumed to be constant across countries in early studies. See Acemoglu et al. (2008) and BM (2015) for detailed descriptions of the variables $d_{it}$ and $I_{it}$.

We use the publicly available data that are used in BM (2015).[7] Following these authors, we consider 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.

### 5.2 Estimation results

First, we can estimate the model in (5.1) by minimizing the non-penalized objective function in (3.4) and ignoring the latent group structure. 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 5. From this figure we see that these estimates have wide dispersion over the plane from which it is hard to discern any pattern.

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 $\eta = 5\%$ of the countries and apply the Information Criterion to choose the tuning parameter as in the simulations. The classification results are displayed in Figure 6, where we use red circle, blue star, green triangle, and black plus to denote Groups 1, 2, 3, and 4, respectively. [See Section G.2 in Wang, Phillips, and Su (2018) for the detailed country-group table.] Interestingly, these four groups distribute in roughly four different quadrants in the plane.

Table 2 reports the estimation results for each group-specific parameter and those for the pooled estimates, all of which are bias-corrected by using Lu and Su's (2017) bias correction formula and Arellano's (1987) country cluster-robust standard errors. The last column in Table 2 reports the estimate of the Long Run Effect of income on democracy, $\beta_1/(1 - \beta_2)$. We summarize some important findings from Table 2. First, Panel-CARDS discovers four latent groups: Group 1 has negative but insignificant $\beta_1$ and positive $\beta_2$; Group 2 has negative $\beta_1$ and negative $\beta_2$; Group 3 has negative $\beta_1$ and positive $\beta_2$; Group 4 has positive $\beta_1$ and negative but insignificant $\beta_2$. These results are consistent with the scatter plot of the preliminary estimates in Figure 6 and suggest the

[7] All the data are directly from AJRY: http://economics.mit.edu/faculty/acemoglu/data/ajry2008.

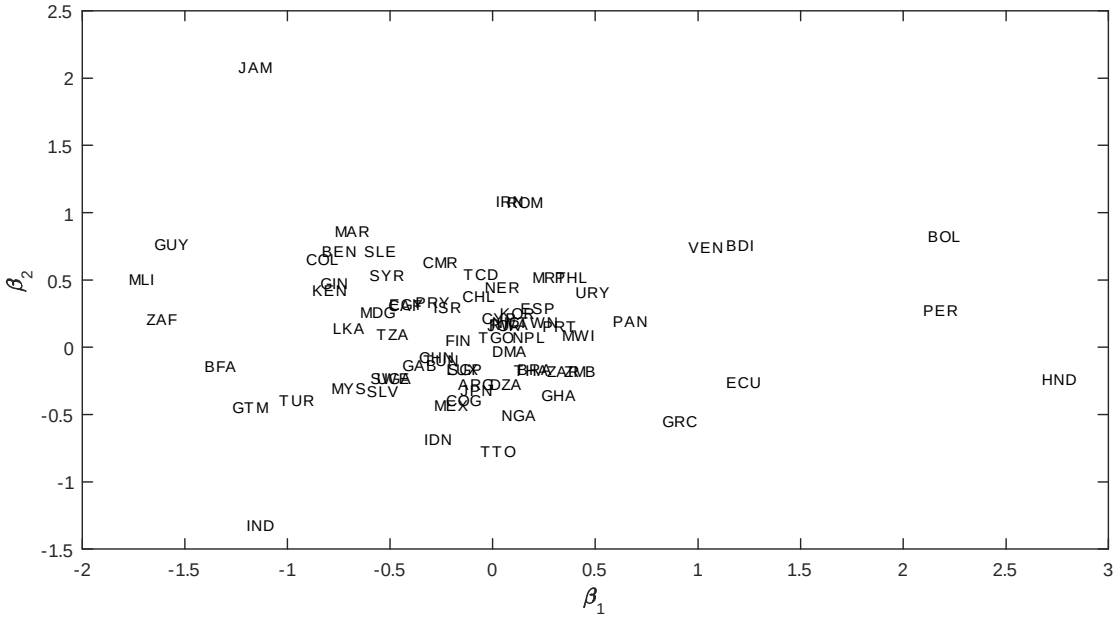


Figure 5: Scatter plot of preliminary estimates. See Table 10 in the online supplement for full names of these 3-letter country codes.

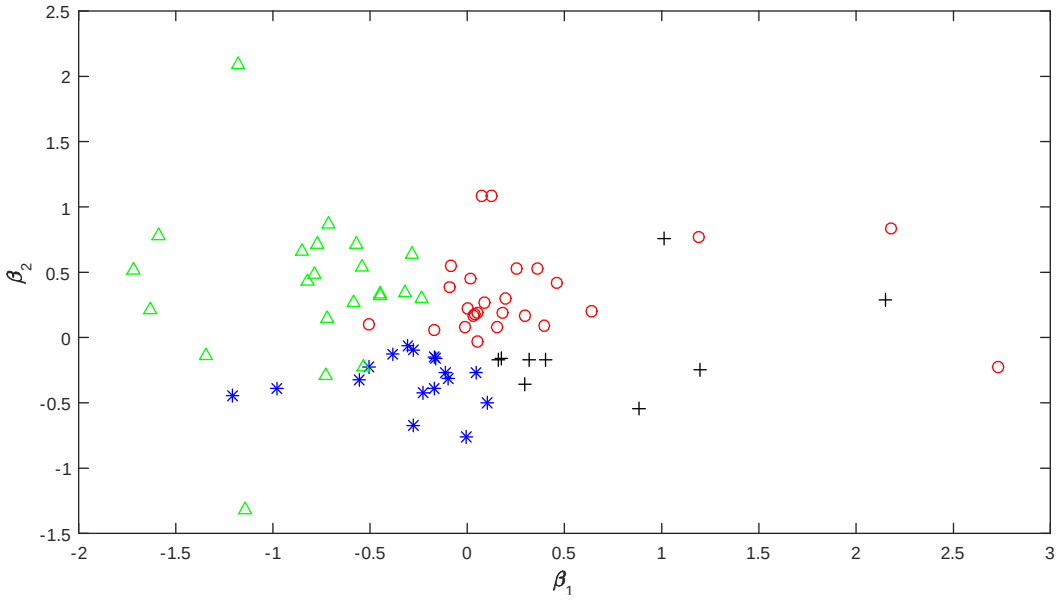


Figure 6: Scatter plot of classification results by using Panel-CARDS. As in Figure 5, the location of the scatter points indicate the value of preliminary estimates of $\beta_1$ and $\beta_2$. The red circle, blue star, green triangle, and black plus correspond to Group 1, Group 2, Group 3, and Group 4, respectively.

Table 2: Regression results for groups 1–4 and the pooled one.

| | $\beta_1$ | | | $\beta_2$ | | | LRE |
|---|---|---|---|---|---|---|---|
| | estimates | s.e. | t-stat | estimates | s.e. | t-stat | |
| Group 1 | 0.024 | 0.024 | 1.005 | 0.364 | 0.081 | 4.481 | 0.038 |
| Group 2 | -0.243 | 0.045 | -5.427 | -0.282 | 0.079 | -3.552 | -0.189 |
| Group 3 | -0.525 | 0.054 | -9.646 | 0.468 | 0.069 | 6.812 | -0.986 |
| Group 4 | 0.380 | 0.093 | 4.071 | -0.149 | 0.132 | -1.127 | 0.331 |
| Pooled FE model | 0.021 | 0.022 | 0.988 | 0.282 | 0.057 | 4.937 | 0.030 |

*Note:* LRE is the abbreviation for the long run effect, which is defined as $\beta_1/(1-\beta_2)$.

effect of income on the level of democracy is not necessarily positive. Second, if we ignore the slope heterogeneity and pool all countries together to estimate a homogeneous panel, the last row of Table 2 indicates a small positive but insignificant effect of income on democracy. Of course, such a regression output cannot explain the observed disparate country-specific income and democracy relationships discussed at the beginning of this section. Third, the estimates of the Long Run Effect for the four groups are 0.038, -0.189, -0.986, and 0.331, which imply that a 10% increase in income per capita is associated with increases of 0.004, -0.019, -0.099, and 0.033 in democracy, respectively. This evidence suggests that income level may have a negative impact on democracy for countries in Group 3, a finding that is at substantial variance with the positive effect from the pooled fixed effect specification that ignores heterogeneity.

# 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 (2015). 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 a longstanding example of panel data research in economics – the international relationship between income and democracy. The methods identify four latent groups of countries which demonstrate distinctive association effects, each relating income to democracy in a different way, some positive and some negative. The application demonstrates that it is possible to take advantage of increased precision in estimation from cross section averaging by identifying those subgroups of a panel in which homogeneous responses are validated by the data while at the same time accommodating heterogeneous responses across groups.


## ACKNOWLEDGMENTS

We would like to thank the co-editor Thierry Magnac and three anonymous referees for their many constructive comments on the paper. Phillips acknowledges support from the NSF (USA) under Grant SES 12-58258 and Grant NRF-2014S1A2A2027803 from the Korean Government. 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.


## References


Acemoglu, D., Johnson, S., Robinson, J. A., & Yared, P. (2008). Income and democracy. *American Economic Review*, 98(3), 808–842.

Amini, A. A., Chen, A., Bickel, P. J., & Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *Annals of Statistics*, 41(4), 2097–2122.

Ando, T., & Bai, J. (2016). Panel data models with grouped factor structure under unknown group membership. *Journal of Applied Econometrics*, 31(1), 163–191.

Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford bulletin of Economics and Statistics*, 49(4), 431–434.

Barro, R. J. (1999). Determinants of democracy. *Journal of Political Economy*, 107(S6), 158–183.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2), 521–547.

Bester, C. A., & Hansen, C. B. (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1), 197–208.

Bondell, H. D., & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1), 115–123.

Bonhomme, S., & Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3), 1147–1184.

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

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

Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

Card, D., & 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(4), 772–793.

Card, D., & 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(5), 1397–1420.

Dhaene, G., & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies*, 82(3), 991–1030.

Dube, A., Lester, T. W., & Reich, M. (2010). Minimum wage effects across state borders: Estimates using contiguous counties. *Review of Economics and Statistics*, 92(4), 945–964.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456), 1348–1360.

Fan, J., Lv, J., & Qi, L. (2011). Sparse high-dimensional models in economics. *Annual Review of Economics*, 3(1), 291–317.

Hsiao, C., & Tahmiscioglu, A. K. (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association*, 92(438), 455–465.

Kasahara, H., & Shimotsu, K. (2009). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica*, 77(1), 135–175.

Ke, Z. T., Fan, J., & Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509), 175–194.

Leeb, H., & Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, 21(1), 21–59.

Leeb, H., & Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics*, 142(1), 201–211.

Pötscher, B. M., & Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100(9), 2065–2082.

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

Lu, X., & Su, L. (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8(3), 729–760.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York: Oxford University Press.

Neumark, D., & Wascher, W. (1992). Employment effects of minimum and subminimum wages: panel data on state minimum wage laws. *Industrial and Labor Relations Review*, 46(1), 55–81.

Neumark, D., & 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(5), 1362–1396.

Neumark, D., & 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., & Tibshirani, R. (2007). Averaged gene expressions for regression. *Biostatistics*, 8(2), 212–227.

Phillips, P. C. B., & Sul, D. (2007). Transition modeling and econometric convergence tests. *Econometrica*, 75(6), 1771–1855.

Sarafidis, V., & Weber, N. (2015). A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*, 77(2), 274–296.

Pötscher, B. M., & Schneider, U. (2009). On the distribution of the adaptive LASSO estimator. *Journal of Statistical Planning and Inference*, 139(8), 2775–2790.

Shen, X., & Huang, H. C. (2010). Grouping pursuit through a regularization solution surface. *Journal of the American Statistical Association*, 105(490), 727–739.

Su, L., & Chen, Q. (2013). Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory*, 29(6), 1079–1135.

Su, L., & Ju, G. (2018). Identifying latent grouped patterns in panel data models with interactive fixed effects. *Journal of Econometrics*, forthcoming.

Su, L., Shi, Z., & Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6), 2215–2264.

Su, L., Wang, X., & Jin, S. (2018). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, forthcoming.

Sun, Y. (2005). Estimation and inference in panel structure models. *Working Paper*, Department 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., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society, Series B*, 67(1), 91–108.

Wang, W., Phillips, P. C. B., & Su, L. (2018). Supplement to "Homogeneity pursuit in panel data models: theory and application". *Journal of Applied Econometrics Supplementary Material*.

# Supplementary Material for "Homogeneity Pursuit in Panel Data Model: Theory and Application"

**(Not for publication)**

Wuyi Wang[a], Peter C.B. Phillips[b], Liangjun Su[c]

[a] Institute for Economic and Social Research, Jinan University

[b] Yale University, University of Auckland, University of Southampton, & Singapore Management University

[c] School of Economics, Singapore Management University

This supplement is composed of seven parts. Section A contains the proofs of the main results in the paper. Section B proves a technical lemma that is used in the proofs of the main results. Sections C and D justify the convergence of the Local Linear Approximation algorithm and the asymptotic validity of the information criterion, respectively. Section E gives more explanations on the construction of the Panel-CARDS objective function. Sections F and G report some additional results on the simulations and application, respectively.

## A Proofs of the Main Results

This section provides the proofs of the main results in the above paper. We will need to refer to Lemma B.1 that is stated and proved in the next section. Throughout we use $M$ to denote a generic positive constant that may vary across lines.

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

**Lemma A.1** *Suppose that Assumption A1 holds. Then for each* $k = 1, \ldots, K$,

*(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 KFW (2015) 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_N')'$ 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 $k = 1, 2, \ldots, K$ and $l = 1, 2, \ldots, L$. For any $\boldsymbol{\beta} \in \mathcal{W}_{NT}^0$, denote $\boldsymbol{\alpha} = S^*(\boldsymbol{\beta})$. Define $n_{km}^{(1)} = \sum_{l=1}^{L-1}(|V_{kl}||V_{m(l+1)}| + |V_{ml}||V_{k(l+1)}|)$,[8] 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* $\boldsymbol{\beta} \neq \boldsymbol{\beta}^*$.

[8] 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 $\frac{t_{NT}}{N N_{\min}} \ll \lambda_2$ 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

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

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

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

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 $\mathbf{z} = (\mathbf{z}_\mathbf{1}', \ldots, \mathbf{z}_\mathbf{N}')'$, $\boldsymbol{\theta}_{ii'}(\mathbf{z}) = \frac{1}{2N_k}(\mathbf{z}_i - \mathbf{z}_{i'})$, 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

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

where the second summation is a telescopic summation by construction 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

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

where

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

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

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

In case (b),

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

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

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

In case (c)

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

It follows that

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

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

$$\mathcal{E}_{31} = \left\{\frac{1}{NT}\max_{1\le k\le K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\theta}_{ij}(\mathbf{z})\right\|_1 \le \frac{M}{NN_{\min}}\left(\frac{\ln T}{\sqrt{T}} + t_{NT}\right)\right\}. \tag{A.22}$$

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 $i$th 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

$$\mathcal{E}_{32}=\left\{\frac{1}{NT}\max_{1\leq k\leq K}\max_{i,j\in G_k^0}\left\|\boldsymbol{\tau}_{ij}(\mathbf{z})\right\|_1\leq\frac{M}{N}\left(\frac{\ln T}{T^{1/2}}+t_{NT}\right)\sqrt{\max_{1\leq k\leq K}\phi_k}\right\}. \tag{A.23}$$

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 $\lambda_1$, $\lambda_2$, and $\phi_k$, 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 $P(\hat{G}_1=G_1^0,\ldots,\hat{G}_K=G_K^0|\hat{K}=K)\to 1$ as $T\to\infty$, perhaps after suitable relabeling among the $G_k^0$'s. In addition,

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

(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)

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

That is, $\sqrt{N_k T}(\hat{\boldsymbol{\alpha}}_{\hat{G}_k} - \boldsymbol{\alpha}_k^0)$ is asymptotically equivalent to $\sqrt{N_k T}(\hat{\boldsymbol{\alpha}}_k^{oracle} - \boldsymbol{\alpha}_k^0)$ 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, $\{\mathcal{B}_{\iota_1}, \ldots, \mathcal{B}_{\iota_R}\}$ 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 $\mathcal{B}^* = \{V_{1d_1}, \ldots, V_{1u_1}, \ldots, V_{Kd_K}, \ldots, V_{Ku_K}\}$ 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, $P(\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3) \geq 1 - \epsilon_1 - o(K/T)$. ■

**Proof of Theorem 3.5**. First, we consider the case with only one ordered segmentation $\mathcal{B}$, i.e., $R = 1$ and the penalized objective function is $Q_{2,NT}(\boldsymbol{\beta}) = L_{2,NT}(\boldsymbol{\beta}) + P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta})$. We show that with obvious modifications both Propositions A.2 and A.3 continue to hold here. By replacing $Q_{NT}$ with $Q_{2,NT}$, we note that the Proposition A.2 still holds because $\hat{\boldsymbol{\alpha}}^{oracle} = S(\hat{\boldsymbol{\beta}}^{oracle})$ is the unique solution to the minimization problem with the convex objective function

$$L^{\mathcal{G}}_{2,NT}(\boldsymbol{\alpha}) \equiv \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} \sum_{t=1}^{T} \left( \ddot{y}_{it} - \tilde{\mathbf{x}}_{it}' \boldsymbol{\alpha}_k + \frac{1}{N} \sum_{\ell=1}^{K} \sum_{j \in G_k} \tilde{\mathbf{x}}_{jt}' \boldsymbol{\alpha}_\ell \right)^2 .$$

In Proposition A.3, the analysis of $I_2$ is still the same. Now we consider $I_1$ in this new setup. We denote $\ddot{\boldsymbol{y}}_i \equiv (\ddot{y}_{i1}, \ldots, \ddot{y}_{iT})'$, $\bar{\mathbf{x}}_{t,\boldsymbol{\beta}} \equiv \frac{1}{N} \sum_{j=1}^{N} \tilde{\mathbf{x}}_{jt}' \boldsymbol{\beta}_j$, $\bar{\mathbf{x}}_{\boldsymbol{\beta}} \equiv (\bar{\mathbf{x}}_{1,\boldsymbol{\beta}}, \ldots, \bar{\mathbf{x}}_{T,\boldsymbol{\beta}})'$, $\bar{\mathbf{z}}_{\boldsymbol{\beta}} = \frac{1}{N} \sum_{i=1}^{N} (\ddot{\boldsymbol{y}}_i - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_i) = -\frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{x}}_i \boldsymbol{\beta}_i = -\bar{\mathbf{x}}_{\boldsymbol{\beta}}$, and similarly for $\bar{\mathbf{x}}_{t,\boldsymbol{\beta}^*}$, $\bar{\mathbf{x}}_{\boldsymbol{\beta}^*}$, and $\bar{\mathbf{z}}_{\boldsymbol{\beta}^*}$. Then

$$\begin{aligned}
I_1 &= L_{2,NT}(\boldsymbol{\beta}) - L_{2,NT}(\boldsymbol{\beta}^*) \\
&= \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \ddot{y}_{it} - \tilde{\mathbf{x}}_{it}' \boldsymbol{\beta}_i + \frac{1}{N} \sum_{j=1}^{N} \tilde{\mathbf{x}}_{jt}' \boldsymbol{\beta}_j \right)^2 \\
&\quad - \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} \sum_{t=1}^{T} \left( \ddot{y}_{it} - \tilde{\mathbf{x}}_{it}' \boldsymbol{\beta}_k^* + \frac{1}{N} \sum_{\ell=1}^{K} \sum_{j \in G_k} \tilde{\mathbf{x}}_{jt}' \boldsymbol{\beta}_\ell^* \right)^2 \\
&= \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} (\ddot{\boldsymbol{y}}_i - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_i)'(\ddot{\boldsymbol{y}}_i - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_i) - \frac{1}{2NT} \sum_{k=1}^{K} \sum_{i \in G_k^0} (\ddot{\boldsymbol{y}}_i - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_k^*)'(\ddot{\boldsymbol{y}}_i - \tilde{\mathbf{x}}_i \boldsymbol{\beta}_k^*) \\
&\quad + \frac{1}{T} (\bar{\mathbf{z}}_{\boldsymbol{\beta}}' \bar{\mathbf{x}}_{\boldsymbol{\beta}} - \bar{\mathbf{z}}_{\boldsymbol{\beta}^*}' \bar{\mathbf{x}}_{\boldsymbol{\beta}^*}) + \frac{1}{2T} (\bar{\mathbf{x}}_{\boldsymbol{\beta}}' \bar{\mathbf{x}}_{\boldsymbol{\beta}} - \bar{\mathbf{x}}_{\boldsymbol{\beta}^*}' \bar{\mathbf{x}}_{\boldsymbol{\beta}^*}) \\
&= -\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}{2T} \sum_{t=1}^{T} (\bar{\mathbf{x}}_{t,\boldsymbol{\beta}} + \bar{\mathbf{x}}_{t,\boldsymbol{\beta}^*}) \frac{1}{N} \sum_{k=1}^{K} \sum_{i \in G_k^0} \tilde{\mathbf{x}}_{it}' (\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)} + \frac{1}{2} (\bar{\mathbf{x}}_{\boldsymbol{\beta}} + \bar{\mathbf{x}}_{\boldsymbol{\beta}^*}))' \tilde{\mathbf{x}}_i (\boldsymbol{\beta}_i - \boldsymbol{\beta}_k^*).
\end{aligned}$$

Note here we can write $I_1$ as in equation (A.13). It can be similarly evaluated and thus Proposition A.3 also holds in the presence of the time fixed effects. By combining Propositions A.2 and A.3, together with Lu and Su's (2017) Theorem 4.1, we can readily prove the case with only one ordered segmentation $\mathcal{B}$. For the general case, the proof follows from that of Theorem 3.4. ■

# B A Technical Lemma

This section states and proves a technical lemma that is used in the proofs in the last section.

**Lemma B.1** *Let $\xi_{it}$ denote a $d_\xi \times 1$ random vector with mean 0 and $\mathbb{E}\|\xi_{it}\|^q < \infty$ for some $q > 4$. Suppose that $\{\xi_{it}, i = 1, \ldots, N,\ t = 1, \ldots, T\}$ are independent across $i$ and are strong mixing in*

*the time index. Let* $G_1^0, \ldots, G_K^0$ *be defined as in the main text with* $N_k = \left|G_k^0\right|$ *for* $k = 1, \ldots, K$. *Let* $\alpha_i(\cdot)$ *denote the mixing coefficients of* $\{\xi_{it}, t = 1, 2, \ldots\}$. *Suppose that* $\alpha_i(\tau) \leq \alpha(\tau)$ *for all* $i = 1, \ldots, N$ *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* $k = 1, \ldots, K$,

*(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 $\iota_\xi$ 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

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

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

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

To prove (i1), we need to rewrite the expression $Q_{1NT} \equiv \frac{1}{N_k T} \sum_{i \in G_k^0} \sum_{t=1}^{T} \xi_{1it}$. Without loss of generality, we assume that we can split the time interval $[1, T]$ into $2r_{N_k T}$ blocks with each block of length $l_{N_k T} = T/\left(2r_{N_k T}\right) \asymp \left(N_k T\right)^{\frac{1}{2}-\vartheta}$ 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

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

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

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

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

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

By the Davydov inequality, we can readily show that

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

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 $B_{i,1}^*$, $B_{i,3}^*, \ldots$ 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 $C_2 \in (0, C_{\xi N_k T}]$,

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

Then we have

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

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 $\mathbb{E}[B_{i,2s-1}] = 0$, we have for $\lambda_{N_kT} \equiv C^{-1}_{\xi N_kT}/2$,

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

Then by the Markov inequality, we have

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

where the last line follows because $\lambda^2_{N_kT}/(N_kT) = \left(\frac{N_kT}{4l_{N_kT}\eta_{N_kT}}\right)^2/(N_kT) = \frac{N_kT}{16l^2_{N_kT}\eta^2_{N_kT}} \asymp l^{-2}_{N_kT}\left(N_kT\right)^{1-2\vartheta} \asymp 1$ 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 $\frac{a_{N_kT}}{4N_kr_{N_kT}} \leq C_{\xi N_kT}$

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

where $L$ can be chosen arbitrarily large as $\alpha\left(l_{N_kT}\right)$ decays to zero at the exponential rate and $l_{N_kT} \asymp \left(N_kT\right)^{\frac{1-2\vartheta}{2}}$ diverges to $\infty$ 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 $\iota_\xi$ 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 $\bar{\xi}_{1it} = \iota_\xi' [\xi_{it}\mathbf{1}_{it} - \mathbb{E}(\xi_{it}\mathbf{1}_{it})]$, $\bar{\xi}_{2it} = \iota_\xi'\xi_{it}\bar{\mathbf{1}}_{it}$, and $\bar{\xi}_{3it} = \iota_\xi'\mathbb{E}(\xi_{it}\bar{\mathbf{1}}_{it})$. 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

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

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

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

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 $l_T = T/(2r_T) \asymp T^{\frac{1}{2}-\bar{\vartheta}}$. 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 $s = 1, \ldots, 2r_T$. 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

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

By the Davydov inequality, we can readily show that

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

for some $\bar{C}_1 < \infty$. By Bradley's lemma, we can construct a sequence of random variables $\bar{B}^*_{i,1}$, $\bar{B}^*_{i,3}, \ldots$ 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

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

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 $\bar{\lambda}_T \equiv \bar{C}^{-1}_{\xi T}/2$ and by the Markov inequality, we have

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

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 $\frac{a_T}{4r_T} \leq \bar{C}_{\xi T}$,

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

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 $l_T = T/(2r_T) \asymp T^{1-\bar{\vartheta}-\epsilon}$ where $\epsilon$ 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 $s = 1,\ldots,2r_T$. 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 $\bar{\lambda}_T^2/T = \left(\frac{T}{4l_T\eta_T}\right)^2/T = \frac{T}{16l_T^2\eta_T^2} = O(l_T^2 T^{1-2\vartheta}) = O\left(T^{-1+2\epsilon}\right) = o(1)$ 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). ■

# C The Convergence of the Local Linear Approximation Algorithm

The section justifies the convergence of Local Linear Approximation algorithm used to obtain the numerical solution.

**Theorem C.1** *Suppose that Assumptions A1–A3 hold. Assume that the initial estimate $\hat{\boldsymbol{\beta}}^{initial}$ along with the tuning parameter $\delta$ generates a segmentation $\mathcal{B}$ admissible with the true grouping pattern with probability at least $1-\epsilon_0$. Assume that $\|\hat{\boldsymbol{\beta}}^{initial} - \boldsymbol{\beta}^0\|_\infty \leq \min(\lambda_1,\lambda_2)$. Then with probability at least $1-\epsilon_0 - o(K/T)$, the Local Linear Approximation algorithm yields $\hat{\boldsymbol{\beta}}^{oracle}$, the oracle estimate, as the stable solution.*

**Proof.** We first prove the first iteration yields $\hat{\boldsymbol{\beta}}^{oracle}$ with high probability, and then show $\hat{\boldsymbol{\beta}}^{oracle}$ is a stable solution. As in the proof of Theorem 3.1, let $\mathcal{E}_1$ denote the segmentation $\mathcal{B}$ that is admissible with the true parameter $\boldsymbol{\beta}^0$, let $\mathcal{E}_2$ be as defined immediately below (A.5), and let $\mathcal{E}_3 = \mathcal{E}_{31} \cap \mathcal{E}_{32}$, where $\mathcal{E}_{31}$ and $\mathcal{E}_{32}$ are defined in (A.22) and (A.23), respectively. On the event $\mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$, we show that Local Linear Approximation gives $\hat{\boldsymbol{\beta}}^{oracle}$ as a stable solution.

Let $v_{ij} = p'_{\lambda_1}(\|\hat{\boldsymbol{\beta}}_i^{initial} - \hat{\boldsymbol{\beta}}_j^{initial}\|_1)$ and $w_{ij} = p'_{\lambda_2}(\|\hat{\boldsymbol{\beta}}_i^{initial} - \hat{\boldsymbol{\beta}}_j^{initial}\|_1)$. In the first iteration, we minimize

$$Q_{NT}^{initial}(\boldsymbol{\beta}) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it}-\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i)^2+\lambda_1\sum_{l=1}^{L-1}\sum_{i\in B_l, j\in B_{l+1}} v_{ij}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_1+\lambda_2\sum_{l=1}^{L}\sum_{i\in B_l, j\in B_l} w_{ij}\|\boldsymbol{\beta}_i-\boldsymbol{\beta}_j\|_1,$$

which is a convex function. Remember $\boldsymbol{\beta}^* = S^{-1}\circ S^*(\boldsymbol{\beta})$, which is defined in the proof of Theorem 3.1. We will show that

$$Q_{NT}^{initial}(\boldsymbol{\beta}) \geq Q_{NT}^{initial}(\boldsymbol{\beta}^*) \geq Q_{NT}^{initial}(\hat{\boldsymbol{\beta}}^{oracle}) \tag{C.1}$$

for any $\boldsymbol{\beta} \in \mathcal{W}_{NT}^0$, where $\mathcal{W}_{NT}^0$ is defined in (A.5).

First, we show the second inequality in (C.1). For $i$ and $j$ in different groups, $\|\boldsymbol{\beta}_i^0 - \boldsymbol{\beta}_j^0\| > 2b_{NT}$. By Assumption A3(ii) and $\|\hat{\boldsymbol{\beta}}^{initial} - \boldsymbol{\beta}^0\|_\infty \leq \lambda_1$, we have $\|\hat{\boldsymbol{\beta}}_i^{initial} - \hat{\boldsymbol{\beta}}_j^{initial}\|_1 \geq 2b_{NT} - 2p\lambda_1 > a\lambda_1$. It follows that $v_{ij} = 0$. Similarly, we can show that for $i$ and $j$ in different groups, $w_{ij} = 0$. For $i$ and $j$ in the same group, $\boldsymbol{\beta}_i = \boldsymbol{\beta}_j$ for $\boldsymbol{\beta} \in M_{\mathcal{G}}$, where $M_{\mathcal{G}}$ is defined in (A.1). So $Q_{NT}^{initial}(\boldsymbol{\beta}^*)$ is reduced to $L_{NT}(\boldsymbol{\beta}^*) = \frac{1}{2NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i^*)^2$, which is convex and takes the unique minimum value when $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}^{oracle}$. Thus the second inequality in (C.1) is proved.

Next, we prove the first inequality in (C.1). By Lagrange mean value theorem and the fact that $Q_{NT}^{initial}(\boldsymbol{\beta}^*) = L_{NT}(\boldsymbol{\beta}^*)$, we have

$$\begin{aligned} Q_{NT}^{initial}(\boldsymbol{\beta}) - Q_{NT}^{initial}(\boldsymbol{\beta}^*) &= -\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i^m)\tilde{\mathbf{x}}_{it}'(\boldsymbol{\beta}_i - \boldsymbol{\beta}_i^*) \\ &\quad +\lambda_1\sum_{l=1}^{L-1}\sum_{i\in B_l, j\in B_{l+1}} v_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1 + \lambda_2\sum_{l=1}^{L}\sum_{i\in B_l, j\in B_l} w_{ij}\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1 \\ &\equiv I_1 + I_2 + I_3. \end{aligned}$$

We only need to show that $I_2 + I_3 > |I_1|$ with probability $1 - o(K/T)$. But this proof is similar to the proof of (A.7). Thus we have shown the first iteration will give $\hat{\boldsymbol{\beta}}^{oracle}$ as the unique solution with probability at least $1 - \epsilon_0 - o(K/T)$.

Now we show the solution $\hat{\boldsymbol{\beta}}^{oracle}$ is stable, i.e., the second iteration still gives $\hat{\boldsymbol{\beta}}^{oracle}$. Note that $\hat{\boldsymbol{\beta}}^{oracle}$ also satisfies our requirements on the $\hat{\boldsymbol{\beta}}^{initial}$ over $\mathcal{E}_1$. By the proof for the first iteration, we see that $\hat{\boldsymbol{\beta}}^{oracle}$ is the unique solution to the optimization of $Q_{NT}^{initial}(\boldsymbol{\beta})$ with $\hat{\boldsymbol{\beta}}^{initial} = \hat{\boldsymbol{\beta}}^{oracle}$. ■

# D Asymptotic Validity of the Information Criterion

**Theorem D.1** *Suppose Assumptions A1–A3 and A5 hold. Then there exists a tuning parameter vector $\boldsymbol{\lambda} = (\delta, \lambda_1, \lambda_2)'$ that satisfies all the requirements of Theorem 3.1. In addition, the information criterion in (2.9) will select a tuning parameter vector that yields the oracle estimator $\hat{\boldsymbol{\beta}}^{oracle}$ with probability approaching 1.*

**Proof.** For a tuning parameters vector $\boldsymbol{\lambda}$ that satisfies all the requirements of Theorem 3.1 and the corresponding Panel-CARDS estimator $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ and mean square error $\hat{\sigma}^2_{\hat{\mathcal{G}}(\boldsymbol{\lambda})}$, we have

$$\begin{aligned}\hat{\sigma}^2_{NT}(\boldsymbol{\lambda}) &= 2L_{NT}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))[\mathbf{1}\{\hat{\mathcal{G}}(\boldsymbol{\lambda}) = \mathcal{G}\} + \mathbf{1}\{\hat{\mathcal{G}}(\boldsymbol{\lambda}) \neq \mathcal{G}\}] \\ &= 2L_{NT}(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}))\mathbf{1}\{\hat{\mathcal{G}}(\boldsymbol{\lambda}) = \mathcal{G}\} + o_P(1) \xrightarrow{P} \sigma_0^2 \text{ as } (N,T) \to \infty.\end{aligned}$$

Then $\text{IC}(\boldsymbol{\lambda}) = \ln(\hat{\sigma}^2_{NT}(\boldsymbol{\lambda})) + pK \cdot \rho_{NT} \to \ln(\sigma_0^2)$ by Assumption A5(ii). For any other $\boldsymbol{\lambda}^*$ which yields $\hat{K}^* = \hat{K}(\boldsymbol{\lambda}^*)$ with $1 \leq \hat{K}^* < K$, by Assumption A5(i) we have

$$\begin{aligned}\text{IC}(\boldsymbol{\lambda}^*) &= \ln(\hat{\sigma}^2_{\hat{\mathcal{G}}_{\hat{K}^*}}) + p\hat{K}^* \cdot \rho_{NT} \geq \min_{1 \leq K^* < K} \min_{\hat{G}_{K^*}} \hat{\sigma}^2_{\hat{G}_{K^*}} + pK^* \cdot \rho_{NT} \\ &\to \bar{\sigma}^2 > \sigma_0^2 \text{ as } (N,T) \to \infty.\end{aligned}$$

Now, we consider any other $\boldsymbol{\lambda}^*$ which yields $\hat{K}^* = \hat{K}(\boldsymbol{\lambda}^*) > K$. Without loss of generality, we assume that $\hat{\mathcal{G}}(\boldsymbol{\lambda}^*)$ is a refinement of the true group structure $\mathcal{G}$. Following the analysis of Lemma S1.14 in SSP (2016), we have $NT[\ln(\hat{\sigma}^2_{\mathcal{G}}) - \ln(\hat{\sigma}^2_{\hat{\mathcal{G}}_{\hat{K}^*}})] = O_P(1)$. Then by Assumption A5(ii)

$$\begin{aligned}P(\hat{K}^* > K) &= P(\text{ IC}(\boldsymbol{\lambda}^*) < \text{IC}(\boldsymbol{\lambda})) \\ &= P(NT[\ln(\hat{\sigma}^2_{\mathcal{G}}) - \ln(\hat{\sigma}^2_{\hat{\mathcal{G}}_{\hat{K}^*}})] > (\hat{K}^* - K)NT\rho_{NT}) \\ &\to 0 \text{ as } (N,T) \to \infty.\end{aligned}$$

In summary, our information criterion will select a tuning parameter vector $\boldsymbol{\lambda}$ that satisfies all the requirements of Theorem 3.1 and yields the oracle estimator w.p.a.1. ■

# E Additional Explanations on the Construction of Panel-CARDS

In this section, we give more explanations on the setting up of the segmentation net, the choice of $R$ and other tuning parameters.

The admissible segmentation net notion is introduced to address some special cases that the admissible segmentation proposed in CARDS cannot handle. Consider the example in the paper, there are three groups and the group-specific parameter vectors are

$$\alpha_1 = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \quad \alpha_2 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}, \quad \text{and } \alpha_3 = \begin{pmatrix} 2 \\ 1 \\ 1 \end{pmatrix}.$$

No matter which regressor/slope coefficient is used to construct the ordered segmentation, the original basic CARDS theory cannot work here. For example, if the classification is based on the order statistics of the estimates of the first slope coefficient, we can separate group 3 from groups 1 and 2 but cannot separate members in group 1 from those in group 2 because these two groups share the same value (1) for the first slope coefficient. This motivates us to propose the admissible segmentation net. Based on the first regressor, we are able to separate group 3 from the other two groups; and based on the second (third) regressor, we can separate group 2 (3) from the other groups. The use of admissible segmentation net has been justified through the theoretical analysis and reinforced through Monte Carlo simulations.

Unfortunately, there is no perfect way to choose $R$. We can think of two extreme cases easily by choosing $R = 1$ or $p$. But as the above discussion suggests, the choice of $R = 1$ may fail to separate one group from the others. On the other hand, the use of $R = p$ is typically unnecessary for large $p$ and it slows down the estimation procedure. Based on our experience, a choice of $R = 2$ typically works very well in both simulations and applications. So in the paper, we offer two practical guidelines to choose the $R < p$ regressors, based on which the segmentations are generated. Like all other Lasso-type methods, the choice of tuning parameters like $\lambda_1$ and $\lambda_2$ rests on the magnitude of the slope coefficients. In practice, we recommend choosing them to minimize the proposed information criterion in (2.9).

# F Additional Results on the Simulations

In this section, we include some additional results on the simulations.

## F.1 More results for DGPs 1-4

We first report more simulation results for DGPs 1–4 in the paper. Tables 3–6 report the frequency of choosing different numbers of groups for DGPs 1–4. These tables suggest that when we set the tuning parameter $\eta$ to be 10%, the 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 $\eta$ decreases, the Panel-CARDS tends to estimate slightly more groups than the correct number of groups for small values of $T$; but its performance quickly improves as $T$ increases.

## F.2 DGPs with dependent regressors or errors

Now, we consider two new DGPs with dependent regressors or errors to check the robustness of the Panel-CARDS.

DGP 5 is the same as DGP 1 in the main text except that here $x_{it,j} = 0.5x_{i,t-1,j} + 0.2\mu_i + e_{it,j}$, $j = 1, 2$ and $\varepsilon_{it} = 0.2\varepsilon_{i,t-1} + e_{it}$, where $e_{it}$, $e_{it,1}$, and $e_{it,2}$ are drawn from i.i.d. standard normal distribution. DGP 6 is the same as DGP 4 in the main text except that $x_{it,j} = 0.5x_{i,t-1,j} + 0.2\mu_i + e_{it,j}$, $j = 1, 2$, where $e_{it,1}$ and $e_{it,2}$ are drawn from i.i.d. standard normal distribution. To generate the $T$ periods of observations for individual $i$ in DGP 5 (resp. DGP 6), we first generate $T + 100$ observations with initialization $e_{i0} = e_{it,1} = e_{it,2} = 0$ (resp. $e_{it,1} = e_{it,2} = 0$), and then throw away the first 100 observations. The tuning parameter $\eta$ takes value 0, 2%, and 5%. Apparently, we have serial dependence in both the regressors and error terms in DGP 5 and in the regressors in DGP 6. [We do not allow for serially correlated errors in DGP 6 to avoid endogeneity.] As before, the number of replications is 200.

Tables 7–8 report the classification results for DGPs 5–6. Comparing the results in these two tables with those in Tables 3 and 6, we find that the presence of dependent regressors and errors have some negative impact on the determination of the correct number of groups when $T$ is small (e.g., $T = 10$), but the negative effect becomes ameliorated as $T$ increases.

Table 9 reports the estimation results for the estimates of the second element in the group-specific vectors $(\{\alpha^0_{k,2}\}_{k=1}^K)$ and $\eta = 2\%$ for DGPs 5–6. It reports the correct classification ratio, RMSE, Bias, and 95% coverage probability of Panel-CARDS in Columns 4–7, and the RMSE, Bias, and 95% coverage probability of the oracle ones in Columns 8–10. They show similar patterns as

Table 3: Frequency of obtaining the estimated number of groups in DGP 1 based on Panel-CARDS. The true number of groups 3 is marked in bold.

| $\eta$ | $N$ | $T$ | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 100 | 10 | 0.000 | 0.060 | 0.825 | 0.110 | 0.005 | 0.000 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.055 | 0.845 | 0.095 | 0.005 | 0.000 | 0.000 | 0.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.985 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.05 | 100 | 10 | 0.000 | 0.000 | 0.505 | 0.380 | 0.105 | 0.010 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.990 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.455 | 0.360 | 0.160 | 0.025 | 0.000 | 0.000 |
| | 200 | 20 | 0.000 | 0.020 | 0.935 | 0.020 | 0.025 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.02 | 100 | 10 | 0.000 | 0.000 | 0.375 | 0.265 | 0.240 | 0.065 | 0.045 | 0.010 |
| | 100 | 20 | 0.000 | 0.000 | 0.935 | 0.065 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.040 | 0.350 | 0.165 | 0.130 | 0.120 | 0.070 | 0.125 |
| | 200 | 20 | 0.000 | 0.000 | 0.955 | 0.015 | 0.010 | 0.005 | 0.000 | 0.015 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0 | 100 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.015 | 0.045 | 0.045 | 0.895 |
| | 100 | 20 | 0.000 | 0.000 | 0.080 | 0.250 | 0.230 | 0.130 | 0.115 | 0.195 |
| | 100 | 40 | 0.000 | 0.000 | 0.705 | 0.270 | 0.020 | 0.005 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 0.985 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.015 | 0.010 | 0.020 | 0.010 | 0.070 | 0.875 |
| | 200 | 20 | 0.000 | 0.025 | 0.050 | 0.060 | 0.135 | 0.175 | 0.135 | 0.420 |
| | 200 | 40 | 0.000 | 0.005 | 0.690 | 0.235 | 0.060 | 0.005 | 0.005 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 4: Frequency of obtaining the estimated number of groups in DGP 2 based on Panel-CARDS. The true number of groups 3 is marked in bold.

| $\eta$ | $N$ | $T$ | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 100 | 10 | 0.000 | 0.060 | 0.780 | 0.160 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.005 | 0.885 | 0.075 | 0.035 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.250 | 0.520 | 0.205 | 0.025 | 0.000 | 0.000 | 0.000 |
| | 200 | 20 | 0.000 | 0.150 | 0.535 | 0.295 | 0.005 | 0.015 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.005 | 0.965 | 0.030 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.05 | 100 | 10 | 0.000 | 0.025 | 0.390 | 0.450 | 0.110 | 0.020 | 0.005 | 0.000 |
| | 100 | 20 | 0.000 | 0.025 | 0.845 | 0.075 | 0.030 | 0.010 | 0.015 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 0.985 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.075 | 0.545 | 0.185 | 0.130 | 0.055 | 0.010 | 0.000 |
| | 200 | 20 | 0.000 | 0.150 | 0.425 | 0.135 | 0.130 | 0.075 | 0.035 | 0.050 |
| | 200 | 40 | 0.000 | 0.000 | 0.825 | 0.070 | 0.040 | 0.060 | 0.005 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.02 | 100 | 10 | 0.000 | 0.005 | 0.340 | 0.375 | 0.195 | 0.080 | 0.005 | 0.000 |
| | 100 | 20 | 0.000 | 0.015 | 0.775 | 0.195 | 0.010 | 0.005 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 0.990 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.050 | 0.235 | 0.355 | 0.155 | 0.095 | 0.065 | 0.045 |
| | 200 | 20 | 0.000 | 0.140 | 0.435 | 0.160 | 0.030 | 0.075 | 0.080 | 0.080 |
| | 200 | 40 | 0.000 | 0.000 | 0.885 | 0.065 | 0.020 | 0.005 | 0.015 | 0.010 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0 | 100 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.015 | 0.010 | 0.065 | 0.910 |
| | 100 | 20 | 0.000 | 0.020 | 0.080 | 0.095 | 0.165 | 0.165 | 0.140 | 0.335 |
| | 100 | 40 | 0.000 | 0.000 | 0.500 | 0.345 | 0.105 | 0.020 | 0.030 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 0.990 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.010 | 0.985 |
| | 200 | 20 | 0.000 | 0.000 | 0.065 | 0.050 | 0.070 | 0.040 | 0.120 | 0.655 |
| | 200 | 40 | 0.000 | 0.000 | 0.325 | 0.305 | 0.125 | 0.090 | 0.070 | 0.085 |
| | 200 | 80 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 5: Frequency of obtaining the estimated number of groups in DGP 3 based on Panel-CARDS. The true number of groups 8 is marked in bold.

| $\eta$ | $N$ | $T$ | 6 | 7 | **8** | 9 | 10 | 11 | 12 | 13+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 100 | 10 | 0.735 | 0.165 | 0.090 | 0.005 | 0.000 | 0.000 | 0.000 | 0.005 |
| | 100 | 20 | 0.000 | 0.000 | 0.930 | 0.070 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.095 | 0.310 | 0.565 | 0.030 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.02 | 100 | 10 | 0.000 | 0.000 | 0.360 | 0.285 | 0.250 | 0.055 | 0.045 | 0.005 |
| | 100 | 20 | 0.000 | 0.000 | 0.925 | 0.075 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.650 | 0.300 | 0.040 | 0.010 | 0.000 | 0.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0 | 100 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.005 | 0.030 | 0.045 | 0.065 | 0.145 | 0.710 |
| | 100 | 40 | 0.000 | 0.000 | 0.520 | 0.280 | 0.165 | 0.035 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.005 | 0.025 | 0.070 | 0.175 | 0.185 | 0.540 |
| | 200 | 40 | 0.000 | 0.000 | 0.690 | 0.265 | 0.020 | 0.025 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 6: Frequency of obtaining the estimated number of groups in DGP 4 based on Panel-CARDS. The true number of groups 3 is marked in bold.

| $\eta$ | $N$ | $T$ | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 100 | 10 | 0.000 | 0.000 | 0.835 | 0.155 | 0.010 | 0.000 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.985 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.815 | 0.145 | 0.035 | 0.005 | 0.000 | 0.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.935 | 0.055 | 0.010 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 0.985 | 0.015 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.02 | 100 | 10 | 0.000 | 0.000 | 0.485 | 0.350 | 0.120 | 0.030 | 0.015 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.935 | 0.030 | 0.025 | 0.010 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.550 | 0.295 | 0.080 | 0.065 | 0.010 | 0.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.875 | 0.075 | 0.020 | 0.025 | 0.005 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0 | 100 | 10 | 0.000 | 0.000 | 0.000 | 0.010 | 0.005 | 0.005 | 0.010 | 0.970 |
| | 100 | 20 | 0.000 | 0.050 | 0.070 | 0.185 | 0.210 | 0.285 | 0.130 | 0.070 |
| | 100 | 40 | 0.000 | 0.005 | 0.930 | 0.060 | 0.005 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.010 | 0.985 |
| | 200 | 20 | 0.000 | 0.010 | 0.040 | 0.120 | 0.100 | 0.150 | 0.090 | 0.490 |
| | 200 | 40 | 0.000 | 0.005 | 0.815 | 0.140 | 0.035 | 0.000 | 0.000 | 0.005 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 7: Frequency of obtaining the estimated number of groups in DGP 5 based on Panel-CARDS. The true number of groups 3 is marked in bold.

| $\eta$ | $N$ | $T$ | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 100 | 10 | 0.000 | 0.000 | 0.420 | 0.355 | 0.145 | 0.080 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.925 | 0.075 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.025 | 0.390 | 0.320 | 0.195 | 0.060 | 0.010 | 0.000 |
| | 200 | 20 | 0.000 | 0.010 | 0.885 | 0.040 | 0.065 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.02 | 100 | 10 | 0.000 | 0.000 | 0.235 | 0.300 | 0.300 | 0.145 | 0.020 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.900 | 0.075 | 0.025 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.270 | 0.145 | 0.160 | 0.120 | 0.085 | 0.220 |
| | 200 | 20 | 0.000 | 0.000 | 0.945 | 0.020 | 0.015 | 0.005 | 0.000 | 0.015 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0 | 100 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.020 | 0.045 | 0.935 |
| | 100 | 20 | 0.000 | 0.000 | 0.120 | 0.115 | 0.225 | 0.245 | 0.120 | 0.175 |
| | 100 | 40 | 0.000 | 0.000 | 0.655 | 0.250 | 0.095 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 0.975 | 0.025 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.000 | 0.005 | 0.000 | 0.005 | 0.035 | 0.955 |
| | 200 | 20 | 0.000 | 0.000 | 0.025 | 0.035 | 0.095 | 0.115 | 0.155 | 0.575 |
| | 200 | 40 | 0.000 | 0.000 | 0.650 | 0.250 | 0.080 | 0.010 | 0.005 | 0.005 |
| | 200 | 80 | 0.000 | 0.000 | 0.990 | 0.010 | 0.000 | 0.000 | 0.000 | 0.000 |

those in DGPs 1 and 4. In particular, as $T$ increases, the performance of Panel-CARDS approaches the oracle estimators fast. This indicates that the Panel-CARDS is robust to DGPs with dependent regressors or error terms.

# G Additional Results on the Application

In this section we include some additional information on the application.

## G.1 Country code and name

For the 74 countries used in the application section, Table 10 reports the country code and country name dictionary.

Table 8: Frequency of obtaining the estimated number of groups in DGP 6 based on Panel-CARDS. The true number of groups 3 is marked in bold.

| $\eta$ | $N$ | $T$ | 1 | 2 | **3** | 4 | 5 | 6 | 7 | 8+ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.05 | 100 | 10 | 0.000 | 0.000 | 0.800 | 0.180 | 0.020 | 0.000 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.865 | 0.105 | 0.010 | 0.010 | 0.010 | 0.000 |
| | 200 | 20 | 0.000 | 0.000 | 0.940 | 0.055 | 0.005 | 0.000 | 0.000 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 0.995 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0.02 | 100 | 10 | 0.000 | 0.000 | 0.510 | 0.285 | 0.190 | 0.015 | 0.000 | 0.000 |
| | 100 | 20 | 0.000 | 0.000 | 0.930 | 0.030 | 0.020 | 0.000 | 0.020 | 0.000 |
| | 100 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.315 | 0.440 | 0.185 | 0.035 | 0.020 | 0.005 |
| | 200 | 20 | 0.000 | 0.000 | 0.925 | 0.000 | 0.040 | 0.005 | 0.030 | 0.000 |
| | 200 | 40 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| 0 | 100 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.025 | 0.975 |
| | 100 | 20 | 0.000 | 0.015 | 0.170 | 0.240 | 0.215 | 0.205 | 0.045 | 0.110 |
| | 100 | 40 | 0.000 | 0.000 | 0.935 | 0.065 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 100 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 10 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 1.000 |
| | 200 | 20 | 0.000 | 0.020 | 0.035 | 0.050 | 0.160 | 0.195 | 0.070 | 0.470 |
| | 200 | 40 | 0.000 | 0.000 | 0.855 | 0.145 | 0.000 | 0.000 | 0.000 | 0.000 |
| | 200 | 80 | 0.000 | 0.000 | 1.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |

Table 9: Correct classification of individuals and point estimation of $\boldsymbol{\alpha}_{\cdot 2}^0$.

| | | | Panel-CARDS | | | | Oracle | | |
|---|---|---|---|---|---|---|---|---|---|
| DGP | $N$ | $T$ | % of Correct Classification | RMSE | Bias | Coverage | RMSE | Bias | Coverage |
| 5 | 100 | 10 | 0.710 | 0.455 | -0.002 | 0.573 | 0.078 | 0.003 | 0.947 |
| | 100 | 20 | 0.946 | 0.200 | 0.005 | 0.837 | 0.056 | 0.004 | 0.953 |
| | 100 | 40 | 0.998 | 0.042 | 0.000 | 0.935 | 0.036 | 0.000 | 0.956 |
| | 100 | 80 | 1 | 0.026 | 0.000 | 0.948 | 0.026 | 0.000 | 0.948 |
| | 200 | 10 | 0.625 | 0.457 | 0.015 | 0.421 | 0.056 | 0.001 | 0.935 |
| | 200 | 20 | 0.917 | 0.208 | 0.006 | 0.809 | 0.038 | -0.000 | 0.952 |
| | 200 | 40 | 0.983 | 0.039 | 0.002 | 0.931 | 0.025 | 0.002 | 0.944 |
| | 200 | 80 | 1 | 0.019 | 0.000 | 0.947 | 0.019 | 0.000 | 0.947 |
| 6 | 100 | 10 | 0.813 | 0.413 | -0.016 | 0.720 | 0.097 | 0.007 | 0.956 |
| | 100 | 20 | 0.973 | 0.150 | 0.009 | 0.895 | 0.059 | 0.008 | 0.942 |
| | 100 | 40 | 0.999 | 0.039 | 0.006 | 0.951 | 0.037 | 0.006 | 0.953 |
| | 100 | 80 | 1 | 0.026 | 0.003 | 0.937 | 0.026 | 0.003 | 0.937 |
| | 200 | 10 | 0.792 | 0.401 | -0.031 | 0.668 | 0.079 | 0.006 | 0.953 |
| | 200 | 20 | 0.964 | 0.156 | 0.004 | 0.890 | 0.045 | -0.007 | 0.946 |
| | 200 | 40 | 0.998 | 0.039 | 0.003 | 0.953 | 0.028 | 0.004 | 0.956 |
| | 200 | 80 | 1 | 0.019 | 0.003 | 0.951 | 0.019 | 0.003 | 0.951 |

Table 11: Classification results of countries/regions. By applying the Panel-CARDS, we get 4 groups.

| | | | | |
|---|---|---|---|---|
| Group 1: "Insignificant $\beta_1$ and positive $\beta_2$" group ($\|\hat{G}_1\| = 26$) | | | | |
| Burundi | Bolivia | Chile | Cyprus | Dominica |
| Spain | Finland | Honduras | Iran | Jordan |
| Korea, Rep. | Mauritania | Malawi | Niger | Nicaragua |
| Nepal | Panama | Philippines | Portugal | Romania |
| Rwanda | Chad | Togo | Taiwan | Tanzania |
| Uruguay | | | | |
| Group 2: "negative $\beta_1$ and negative $\beta_2$" group ($\|\hat{G}_2\| = 17$) | | | | |
| Argentina | China | Congo, Rep. | Algeria | Gabon |
| Guatemala | Indonesia | Japan | Luxembourg | Mexico |
| Nigeria | Singapore | El Salvador | Trinidad and Tobago | Tunisia |
| Turkey | Uganda | | | |
| Group 3: "negative $\beta_1$ and positive $\beta_2$" group ($\|\hat{G}_3\| = 22$) | | | | |
| Benin | Burkina Faso | Central African Republic | Cameroon | Colombia |
| Egypt, Arab Rep. | Guinea | Guyana | India | Israel |
| Jamaica | Kenya | Sri Lanka | Morocco | Madagascar |
| Mali | Malaysia | Paraguay | Sierra Leone | Sweden |
| Syrian Arab Republic | South Africa | | | |
| Group 4: "positive $\beta_1$ and insignificant $\beta_2$" group ($\|\hat{G}_3\| = 9$) | | | | |
| Brazil | Ecuador | Ghana | Greece | Peru |
| Thailand | Venezuela, RB | Congo, Dem. Rep. | Zambia | |

Table 10: Dictionary for country codes and names.

| Code | Name | Code | Name | Code | Name | Code | Name |
|---|---|---|---|---|---|---|---|
| ARG | Argentina | GAB | Gabon | MDG | Madagascar | SLV | El Salvador |
| BDI | Burundi | GHA | Ghana | MEX | Mexico | SWE | Sweden |
| BEN | Benin | GIN | Guinea | MLI | Mali | SYR | Syrian Arab Rep. |
| BFA | Burkina Faso | GRC | Greece | MRT | Mauritania | TCD | Chad |
| BOL | Bolivia | GTM | Guatemala | MWI | Malawi | TGO | Togo |
| BRA | Brazil | GUY | Guyana | MYS | Malaysia | THA | Thailand |
| CAF | Central African Rep. | HND | Honduras | NER | Niger | TTO | Trinidad and Tobago |
| CHL | Chile | IDN | Indonesia | NGA | Nigeria | TUN | Tunisia |
| CHN | China | IND | India | NIC | Nicaragua | TUR | Turkey |
| CMR | Cameroon | IRN | Iran | NPL | Nepal | TWN | Taiwan |
| COG | Congo, Rep. | ISR | Israel | PAN | Panama | TZA | Tanzania |
| COL | Colombia | JAM | Jamaica | PER | Peru | UGA | Uganda |
| CYP | Cyprus | JOR | Jordan | PHL | Philippines | URY | Uruguay |
| DMA | Dominica | JPN | Japan | PRT | Portugal | VEN | Venezuela, RB |
| DZA | Algeria | KEN | Kenya | PRY | Paraguay | ZAF | South Africa |
| ECU | Ecuador | KOR | Korea, Rep. | ROM | Romania | ZAR | Congo, Dem. Rep. |
| EGY | Egypt | LKA | Sri Lanka | RWA | Rwanda | ZMB | Zambia |
| ESP | Spain | LUX | Luxembourg | SGP | Singapore | | |
| FIN | Finland | MAR | Morocco | SLE | Sierra Leone | | |

## G.2 Countries within each estimated group

Table 11 reports the countries within each of the four estimated groups. It suggests that each group contains a fairly large number of countries.

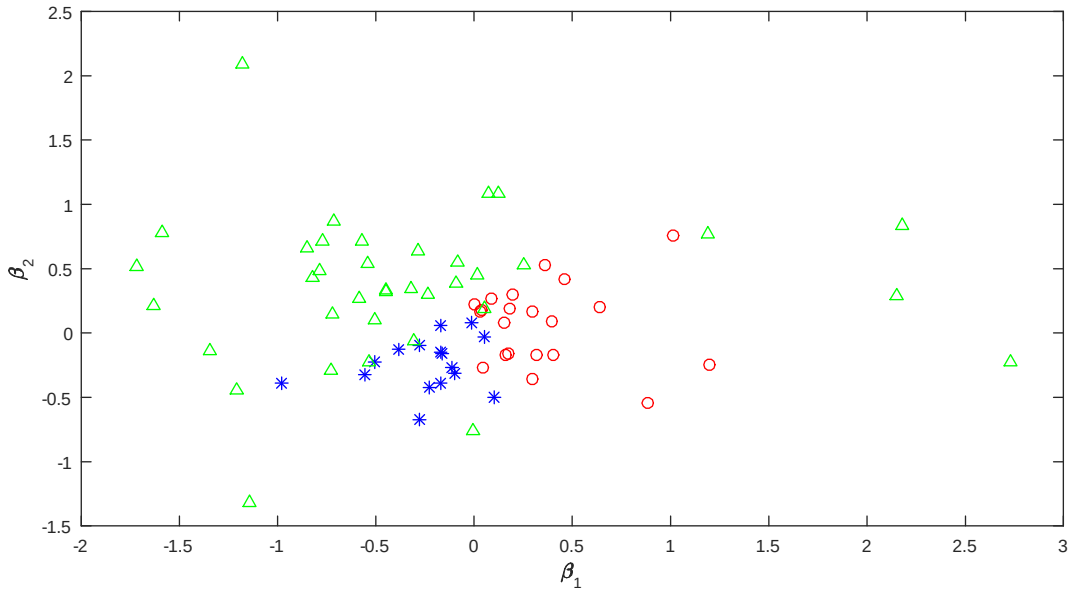


Figure 7: Scatter plot of classification results by Lu and Su (2017). The points indicate the value of the preliminary estimates of $\beta_1$ and $\beta_2$. The red circle, blue star, and green triangle correspond to Groups 1, 2, and 3, respectively.

## G.3 Comparison with the results in Lu and Su (2017)

Lu and Su (2017) also study the estimation of the model in (5.1) by using SSP's (2016) C-Lasso method. So it is worthwhile to compare their results with ours.

Interestingly, Lu and Su (2017) identify three groups while our Panel-CARDS finds four groups. Figure 7 displays the Lu and Su's (2017) classification results based on the C-Lasso method. A close comparison of regression outputs with those in Lu and Su (2017) suggests that our Groups 2 and 3 have similar estimates of $\beta_1$ and $\beta_2$ as those for their Groups 2 and Group 3, respectively. The major difference lies between our Groups 1 and 4 and their Group 1. Despite this, we find the Normalized Mutual Information between the two sets of estimated group structures is 0.4241, which indicates that they overlap with each other surprisingly well.[9]

### References

Acemoglu, D., Johnson, S., Robinson, J. A., & Yared, P. (2008). Income and democracy. *American Economic Review*, 98(3), 808–842.

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

Dhaene, G., & Jochmans, K. (2015). Split-panel jackknife estimation of fixed-effect models. *Review of Economic Studies*, 82(3), 991–1030.

Ke, Z. T., Fan, J., & Wu, Y. (2015). Homogeneity pursuit. *Journal of the American Statistical Association*, 110(509), 175–194.

[9] Amini et al. (2013) give some benchmark values of NMI for reference: for large $N$, matching 50%, 70%, and 90% of the labels correspond to values of NMI of approximately 0.12, 0.26, and 0.58, respectively.

Lu, X., & Su, L. (2017). Determining the number of groups in latent panel structures with an application to income and democracy. *Quantitative Economics*, 8(3), 729–760.

Su, L., Shi, Z., & Phillips, P. C. B. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6), 2215–2264.