Received: 30 November 2016 | Revised: 26 February 2018

DOI: 10.1002/jae.2632

RESEARCH ARTICLE

WILEY Journal of APPLIED ECONOMETRICS

Homogeneity pursuit in panel data models: Theory and application

Wuyi Wang[1] | Peter C. B. Phillips[2,3,4,5] | Liangjun Su[5]

[1]Institute for Economic and Social Research, Jinan University, Guangzhou, China

[2]Department of Economics, Yale University, New Haven, Connecticut

[3]Department of Economics, University of Auckland, Auckland, New Zealand

[4]Department of Economics, University of Southampton, Southampton, UK

[5]School of Economics, Singapore Management University, Singapore

Correspondence
Liangjun Su, School of Economics, Singapore Management University, 90 Stamford Road, Singapore 178903.
Email: ljsu@smu.edu.sg

Funding information
NSF (USA), Grant/Award Number: SES 12-58258 and NRF-2014S1A2A2027803; Singapore Ministry of Education, Grant/Award Number: MOE2012-T2-2-021

Summary

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.

1 | INTRODUCTION

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

This paper follows earlier work by Su, Shi, and Phillips (2016) 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 & Sul, 2007); in financial markets, stock returns in the same sector are commonly thought to share common characteristics (Ke, Fan, & Wu, 2015); and in economic geography, location may be a relevant factor in economic performance, leading to spatial geographic groupings in the data (Bester & Hansen, 2016; Fan, Lv, & Qi, 2011).

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 & 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 & Hansen, 2016). However, 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, Su et al. (2016) developed 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, Su et al. proposed a Bayesian-type information criterion to determine the number of groups. In addition, Lu and Su (2017) proposed a direct testing procedure to identify the number of groups, and Su and Ju (2018) and Su, Wang, and Jin (2018) extended 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 De 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, Saunders, Rosset, Zhu, & Knight, 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 & Reich, 2008; Shen & Huang, 2010).

Ke et al. (2015) explore homogeneity in regressions by designing a method called CARDS (clustering algorithm in regression via data-driven segmentation). They first estimate the parameters by ordinary least squares 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. Ke et al. 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, whereas CARDS does so on individual slope differences. In a panel data model with $p > 1$ regressors, Ke et al.'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 Ke et al. to the segmentation net, which enables us to extract groups more accurately. Just like CARDS for cross-section data or the Su et al. (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

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

here. Second, like Su et al.'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, whereas our Panel-CARDS method does not need to do so. In fact, existing theories for either 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 a 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, Ke et al. (2015) require nonstochastic regressors and sub-Gaussian errors, whereas we permit random regressors or lagged dependent variables, and replace the sub-Gaussian requirement with 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 normal, as discussed in Leeb and Pöscher (2005, 2008), Pötscher and Leeb (2009), and Pötscher and Schneider (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 reinvestigates relationships between income and democracy, a matter that has attracted considerable interest among political economists (cf. Acemoglu, Johnson, Robinson, & Yared, 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. An online Supplement (Supporting Information) 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 $\|\boldsymbol{\alpha}\|_q = \left(\sum_{j=1}^n |\alpha_j|^q\right)^{1/q}$ with $1 \le q < \infty$. When $q = 2$, we abbreviate $\|\cdot\|_2$ as $\|\cdot\|$. Let $\|\boldsymbol{\alpha}\|_\infty = \max_{1\le j\le n}|\alpha_j|$. 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 Su et al. (2016), we consider a panel data model with latent group structure

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

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

$$\boldsymbol{\beta}_i^0 = \sum_{k=1}^{K} \boldsymbol{\alpha}_k^0 \cdot \mathbf{1}\{i \in G_k^0\}. \tag{2}$$

Here $\boldsymbol{\alpha}_l^0 \neq \boldsymbol{\alpha}_k^0$ for any $l \neq k$, and $\mathcal{G} = \{G_1^0, G_2^0, \cdots, G_K^0\}$ 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 $\boldsymbol{\alpha} \equiv (\boldsymbol{\alpha}_1', \cdots, \boldsymbol{\alpha}_K')'$ and $\boldsymbol{\beta} \equiv (\boldsymbol{\beta}_1', \cdots, \boldsymbol{\beta}_N')'$. 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 Equation 2, we can estimate the model (Equation 1) directly. After concentrating out the fixed effects, we obtain the objective function

$$L_{NT}(\boldsymbol{\beta}) = \frac{1}{2NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}' \boldsymbol{\beta}_i)^2, \tag{3}$$

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 ordinary least squares 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 $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \tilde{\boldsymbol{\beta}}_2', \cdots, \tilde{\boldsymbol{\beta}}_N')'$ and $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \cdots, \tilde{\boldsymbol{\beta}}_N)$, 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 Ke et al. (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 $\tilde{\boldsymbol{B}}$, 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 $\imath = 1, 2, \ldots, p$ 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, Whinston, & Green, 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." Thus we have the following theorem.

Theorem 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, that is, one particular row of $\tilde{\mathbf{B}}$.*

Now we only need to select a proper element $\imath^*$ 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 $\imath^*$ denote the index of the regressor that has the largest variation across individuals for its coefficient estimates. Then we can sort $\{\tilde{\beta}_{i\imath^*}, i = 1, 2, \ldots, N\}$ to obtain the order

$$\tilde{\beta}_{\tau(1)\imath^*} \leq \tilde{\beta}_{\tau(2)\imath^*} \leq \cdots \leq \tilde{\beta}_{\tau(N)\imath^*}. \tag{4}$$

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

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

Note that when $p = 1$, an ordered segmentation (Ke et al., 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 Equation 4 together with a tuning parameter δ we have a segmentation comprising $B_1 = \{2,3\}, B_2 = \{1,5\}, B_3 = \{4,6,7\}, B_4 = \{9,10\}$, and $B_5 = \{8\}$, then we can easily verify that the segmentation is admissible by Definition 1.[2] However, the segmentation $\mathcal{B} = \{B_1, \cdots, B_5\}$ with $B_1 = \{2,3\}$, $B_2 = \{1,5,7\}$, $B_3 = \{4,6\}, B_4 = \{9,10\}$ and $B_5 = \{8\}$ is not admissible.

To rank vectors, we need to ensure the admissibility of a segmentation. However, 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 of obtaining an admissible segmentation, no matter how accurate the preliminary estimates are. On the other hand, if we 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

$$(\boldsymbol{\alpha}_1^0, \boldsymbol{\alpha}_2^0, \boldsymbol{\alpha}_3^0) = \left(\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \right), \tag{5}$$

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 $\mathcal{B} = \{B_1, B_2, \cdots, B_L\}$. As in the Ke et al. (2015) CARDS algorithm, we propose the following hybrid penalty:

$$P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \underbrace{\sum_{l=1}^{L-1} \sum_{i \in B_l, j \in B_{l+1}} p_{\lambda_1}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1)}_{\text{between-segment penalty}} + \underbrace{\sum_{l=1}^{L} \sum_{i \in B_l, j \in B_l} p_{\lambda_2}(\|\boldsymbol{\beta}_i - \boldsymbol{\beta}_j\|_1)}_{\text{within-segment penalty}}, \tag{6}$$

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 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 (Equation 6) to the original objective function (Equation 3), we obtain the following penalized least squares objective function:

$$Q_{NT}(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{\mathcal{B},\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{7}$$

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:

$$\hat{\boldsymbol{\beta}}^{(s+1)} = \arg\min \left\{ L_{NT}(\boldsymbol{\beta}) + R(\hat{\boldsymbol{\beta}}^{(s)}; \boldsymbol{\beta}) \right\}, \tag{8}$$

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 Equation 8 is convex, we can apply a standard convex optimization package to obtain the solution.

---

[2]The value of δdetermines 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 $\tilde{\beta}_{1\iota^*} - \tilde{\beta}_{3\iota^*} > \delta$, $\ldots \tilde{\beta}_{8\iota^*} - \tilde{\beta}_{10\iota^*} > \delta$, 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.

The justification of using local linear approximation to solve Equation 7 is relegated to the Supporting Information. We use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda)$ to denote the final solution.

Evidently, the performance of $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\lambda)$ depends on the choice of λ. Following Su et al. (2016), we can choose $\lambda = (\delta, \lambda_1, \lambda_2)'$ to minimize the following information criterion:

$$\mathrm{IC}(\lambda) = \ln\left(\sigma^2_{NT}(\lambda)\right) + pK(\lambda)\rho_{NT}, \tag{9}$$

where $\sigma^2_{NT}(\lambda)$ and $K(\lambda)$ are estimates of σ^2 and the number of groups associated with λ, and $\rho_{NT} = 0.5(NT)^{-1/2}$.[4] In the Supporting Information, we show that 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 Equation 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 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 $\mathcal{G} = \{G_1^0, G_2^0, \cdots, G_K^0\}$ denote the true grouping structure. Given R segmentations $\mathcal{B}_{\iota_1}, \cdots, \mathcal{B}_{\iota_R}$, 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 $\mathcal{N} \equiv \{\mathcal{B}_{\iota_1}, \cdots, \mathcal{B}_{\iota_R}\}$ 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 an appropriate way to derive the true grouping property. Based on this idea, we propose an advanced version of Panel-CARDS that can be regarded as an extension of the above basic Panel-CARDS procedure. Given an admissible segmentation net $\mathcal{N} = \{\mathcal{B}_{\iota_1}, \cdots, \mathcal{B}_{\iota_R}\}$, the advanced Panel-CARDS algorithm is as follows:

- For each $\mathcal{B}_{\iota_r}$, we construct the penalty function $P_{\mathcal{B}_{\iota_r},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ as introduced in Equation 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:

$$Q^*_{NT}(\boldsymbol{\beta}) = L_{NT}(\boldsymbol{\beta}) + P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}). \tag{10}$$

Advanced Panel-CARDS reduces to basic Panel-CARDS in the case that $R = 1$. 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 do not use exhaustive pairwise penalties, which are challenging in computation and not accurate in statistical inference (as in Ke et al., 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 $\{\tilde{\beta}_{1j}, \cdots, \tilde{\beta}_{nj}\}$ 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 that are most likely to have heterogeneous responses. For further explanations, see Section E of the Supporting Information.

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 that cannot be properly segmented by any segmentation), the classification based on the basic Panel-CARDS may still perform reasonably well in finite samples.

[4]Too small or too large a δ will generate too many or too few segments that are not ideal in achieving correct identification. In practice, we find it helpful to set the number of segments directly, which is also easy to control. For example, when $N = 100$, we try $L = 10$, 20, and 30. The choices of λ_1 and λ_2 depend on the value of coefficients we use in the data generating process (DGP). Generally speaking, when the coefficients are large, the tuning parameters λ_1 and λ_2 are large correspondingly.
[5]Its existence follows directly from Theorem 3 of Ke et al. (2015).

2.3 | Hierarchical clustering

When the signal-to-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 misclassification. 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, Hastie, and Tibshirani (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 into the procedure. The hierarchical clustering is used here to improve the finite sample performance, and its asymptotic theory can be justified provided such a priori 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 $\tilde{\mathbf{x}}_i = (\tilde{\mathbf{x}}_{i1}, \cdots, \tilde{\mathbf{x}}_{iT})'$, $\tilde{\mathbf{y}}_i = (\tilde{y}_{i1}, \cdots, \tilde{y}_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})'$, and $\mathbf{y}_i = (y_{i1}, \dots, y_{iT})'$. Let $\max_{i,t}$ denote $\max_{1\le i\le N}\max_{1\le t\le T}$. 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 $p'_{\lambda_j}(s) = \mathrm{d}p_{\lambda_j}(s)/\mathrm{d}s$ for $j = 1, 2$. Let $b_{NT} = \frac{1}{2}\min_{1\le k<j\le K}\|\boldsymbol{\alpha}_k^0 - \boldsymbol{\alpha}_j^0\|_1$. Given $\{G_k^0\}$ and segmentation $\{B_1, \dots, B_L\}$, we define $\Phi_k = N_k/\min\{N_k^3, \min_{d_k\le l\le u_k}|B_l|^2\}$. Note that $1/N_k^2 \le \Phi_k \le N_k$. Let $\hat{G}_{\hat{K}} = \{\hat{G}_1, \hat{G}_2, \cdots, \hat{G}_{\hat{K}}\}$ be an arbitrary partition of $\{1, \dots, N\}$ where $|\hat{G}_k| \ge 1$ for $k = 1, \cdots, \hat{K}$. Define $\hat{\sigma}^2_{\hat{G}_{\hat{K}}} = (NT)^{-1}\sum_{k=1}^{\hat{K}}\sum_{i\in\hat{G}_k}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{x}'_{it}\check{\beta}_i)^2$, where $\{\check{\beta}_i\}$ solves the minimization problem with objective function $L_{NT}(\beta)$ and the constraint imposed by the group structure $\hat{G}_{\hat{K}}$. We use $(N_k, T) \to \infty$ to signify that N_k and T pass to infinity jointly.

We make the following assumptions.

Assumption 1.

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

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

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

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

Assumption 2. $p_\lambda(\cdot)$ 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 $|s| \ge a\lambda$.

Assumption 3.

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

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

Assumption 4.

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

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

Assumption 5.

(i) As $(N, T) \to \infty$, $\min_{1 \le \hat{K} < K} \min_{\hat{G}_{\hat{K}}} \hat{\sigma}^2_{\hat{G}_{\hat{K}}} \xrightarrow{P} \bar{\sigma}^2 > \sigma_0^2$, 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 1(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 Assumption 1(i) does not rule out serial correlation among $\{\epsilon_{it}, t = 1, 2, \cdots\}$ or $\{\mathbf{x}_{it}\epsilon_{it}, t = 1, 2, \cdots\}$. Assumption 1(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. Assumption 1(iii) imposes some moment conditions on $\mathbf{x}_{it}$ and ϵ_{it}. In comparison with conditions 1 and 3 in Ke et al. (2015), which require nonrandom regressors and sub-Gaussian error terms, the conditions in Assumptions 1(i)–(iii) are quite weak. Assumption 1(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, Su et al. (2016) require that N_k passes to infinity at the same rate as N for each k.

Assumption 2 is identical to condition 2 in Ke et al. (2015). Following Ke et al., we specify $p_\lambda(\cdot)$ as the SCAD penalty function in our simulations and the application below. Assumption 3 imposes conditions on K, b_{NT}, λ_1, and λ_2. Assumption 3(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 Su et al. (2016). Assumption 3(ii) specifies the ranges of speed at which λ_1 and λ_2 shrink to zero. Assumption 4 is borrowed from Su et al. 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 Su et al. for detailed discussions concerning Assumption 4. Assumption 5 is imposed to ensure the asymptotic validity of our information criterion (Equation 9). Assumption 5(i) assumes that for all underfitted models the mean square errors would be asymptotically greater than σ_0^2, and Assumption 5(i) is imposed to avoid both over- and underfitted 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 $\mathcal{G} = \{G_1^0, \cdots, G_K^0\}$ 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 Equation 3. The resulting estimator of $\boldsymbol{\beta}$ is denoted by $\hat{\boldsymbol{\beta}}^{\text{oracle}}$. Similarly, by using the true grouping structure, we obtain the oracle estimator $\hat{\boldsymbol{\alpha}}^{\text{oracle}}$ of $\boldsymbol{\alpha}$ with a typical block given by

$$\hat{\boldsymbol{\alpha}}_k^{\text{oracle}} = \left(\sum_{i \in G_k^0} \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i \right)^{-1} \sum_{i \in G_k^0} \tilde{\mathbf{x}}_i' \tilde{\mathbf{y}}_i \quad \text{for } k = 1, \cdots, K. \tag{11}$$

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

Theorem 2. *Suppose that Assumptions 1–3 hold. Suppose that the preliminary estimate $\tilde{\boldsymbol{\beta}}$ and tuning parameter δ together generate a segmentation $\mathcal{B}$ admissible with the true grouping pattern with probability at least $1 - \epsilon_0$. Then with probability at least $1 - \epsilon_0 - o(K/T)$, the Panel-CARDS objective function (Equation 7) has a strictly local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2', \cdots, \hat{\boldsymbol{\beta}}_N')'$ such that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{oracle}$ and $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\| = O_p(\sqrt{K/T})$.*

Theorem 2 parallels Theorem 2 in Ke et al. (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 $P(\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{oracle}}) \to 1$ 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 $\hat{G}_k$, $k = 1, 2, \cdots, \hat{K}$ to denote the $\hat{K}$ estimated groups.

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

$$\hat{G}_k = \left\{ i \in \{1, 2, \cdots, N\} : \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\alpha}}_k \right\} \text{ for } k = 1, 2, \cdots, \hat{K}. \tag{12}$$

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

Theorem 3. *Suppose that the conditions in Theorem 2 are satisfied. Suppose that Assumption 4 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* $P\left(\hat{G}_1 = G_1^0, \cdots, \hat{G}_K = G_K^0\right) \to 1$ *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(i) indicates that w.p.a.1 we can determine the correct number of groups. Theorem 3(ii) reports the asymptotic distribution of the group-specific estimator. As Su et al. (2016) remark, the oracle estimator $\hat{\boldsymbol{\alpha}}_k^{\text{oracle}}$ satisfies

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

under Assumption 4. Theorem 3(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

$$\hat{\boldsymbol{\alpha}}_{\hat{G}_k} = \left(\sum_{i \in \hat{G}_k} \tilde{\mathbf{x}}_i' \tilde{\mathbf{x}}_i\right)^{-1} \sum_{i \in \hat{G}_k} \tilde{\mathbf{x}}_i' \tilde{\mathbf{y}}_i, k = 1, \cdots, \hat{K}. \tag{13}$$

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

Theorem 4. *Suppose that the conditions in Theorem 3 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* $T \to \infty$.

Thus 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 Su et al. (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 5. *Suppose that Assumptions 1–A3 hold. Suppose that the preliminary estimate $\tilde{\boldsymbol{\beta}}$, the tuning parameter δ, and the choice of R together generate an admissible segmentation net $\mathcal{N}$ with probability at least $1 - \epsilon_1$. Then, with*

[6] See Theorem 3 in Ke et al. (2015) for a proof.

probability at least $1-\varepsilon_1-o(K/T)$*, the Panel-CARDS objective function (Equation 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 $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ 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, $\varepsilon_1\equiv\varepsilon_{1T}\to 0$ and we have $P\left(\hat{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}^{\text{oracle}}\right)\to 1$ as $T\to\infty$. Then analogous results as in Theorems 3 and 4 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:

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

where γ_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 Equation 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

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

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

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

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

$$L_{2,NT}(\boldsymbol{\beta})=\frac{1}{2NT}\sum_{i=1}^N\sum_{t=1}^T\left(\ddot{y}_{it}-\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_i+\frac{1}{N}\sum_{j=1}^N\tilde{\mathbf{x}}_{jt}'\boldsymbol{\beta}_j\right)^2. \tag{14}$$

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

$$Q_{2,NT}^*(\boldsymbol{\beta})=L_{2,NT}(\boldsymbol{\beta})+P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta}), \tag{15}$$

where $P_{\mathcal{N},\lambda_1,\lambda_2}(\boldsymbol{\beta})$ is as defined in Equation 10. By solving Equation 15 we obtain the Panel-CARDS estimator $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_1',\hat{\boldsymbol{\beta}}_2',\cdots,\hat{\boldsymbol{\beta}}_N')'$.Let $\hat{\boldsymbol{\beta}}^{\text{oracle}}$ denote the oracle estimator of $\boldsymbol{\beta}$ by knowing the true group structure of $\boldsymbol{\beta}_i$s in Equation 2. Let $\hat{\boldsymbol{\alpha}}^{\text{oracle}}=(\hat{\boldsymbol{\alpha}}_1^{oracle'},\cdots,\hat{\boldsymbol{\alpha}}_K^{\text{oracle}'})'$ denote the group-specific version of $\hat{\boldsymbol{\beta}}^{\text{oracle}}$.

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

Theorem 6. *Suppose that Assumptions 1–3 hold. Suppose that the preliminary estimate* $\tilde{\boldsymbol{\beta}}$*, the tuning parameter* δ*, and the choice of* R *together generate an admissible segmentation net* $\mathcal{N}$ *with probability at least* $1-\varepsilon_1$*. Then with probability at least* $1-\varepsilon_1-o(K/T)$*, the Panel-CARDS objective function in Equation 15 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 $\hat{\boldsymbol{\alpha}}^{\text{oracle}}$; 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 DGPs.

DGP 1. Both the fixed effects μ_i and the error terms follow the i.i.d. standard normal distribution across time, and individuals and are mutually independent of each other. Individuals are divided into three groups with $N_1 : N_2 : N_3 = 4 : 3 : 3$. The observations $(y_{it}, \mathbf{x}_{it})$ are generated from the panel model (Equation 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 Su et al. (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 (Equation 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 three 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 Equation 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 η controls the minimum percentage of observations within each estimated group. We set $\eta = 10\%, 5\%, 2\%$, and 0 to estimate the model and obtain the grouping results. When $\eta = 0$, we allow the minimum number of elements in an estimated group to be 1. The larger the value of η, the larger the number of elements for the smallest estimated group that is allowed and the smaller the number of groups estimated. For DGPs 1 and 2, we consider all candidate values of η : $10\%, 5\%, 2\%$, and 0; for DGPs 3 and4, we consider $\eta = 5\%, 2\%$, and 0 because $\eta = 10\%$ is a strong assumption when we have eight groups in DGP 3.

The hierarchical clustering is carried out as follows.

- Let $N^* = N\eta$. For a Panel-CARDS classification $\mathcal{A}^0 = \{A_1, A_2, \cdots, A_{\hat{K}^0}\}$, 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 $\mathcal{A} = \{A_1, A_2, \cdots, A_{\hat{K}}\}$, 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 rerun the classification.

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

$$k^* = \underset{k\in\{1,2,\cdots,\hat{K}^0\};\beta_1,\cdots,\beta_{\hat{K}}}{\arg\min} \frac{1}{2NT}\sum_{l=1}^{\hat{K}}\sum_{i\in A_l}\sum_{t=1}^{T}\left[(\tilde{y}_{it}-\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}_l)^2+(\tilde{y}_{jt}-\tilde{\mathbf{x}}_{jt}'\boldsymbol{\beta}_k)^2\cdot\mathbf{1}\{k=l\}\right].$$

  Now we reclassify the element j to group A_{k^*} for $k^* \in \{1,\cdots,\hat{K}\}$. In other words, we treat j as a new observation, and reclassify it to the group that 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 $\hat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, \cdots, \hat{G}_{\hat{K}}\}$.

We use a Bayesian-type information criterion to choose the tuning parameters. Given the Panel-CARDS classification results $\hat{\mathcal{G}} = \{\hat{G}_1, \hat{G}_2, \cdots, \hat{G}_{\hat{K}}\}$, which are obtained by using the tuning parameter vector λ, we calculate $\text{IC}(\lambda) = \ln\left(\sigma_{NT}^2(\lambda)\right) + p\hat{K}/(2\sqrt{NT})$, where $\sigma_{NT}^2(\lambda) = \frac{1}{NT}\sum_{s=1}^{\hat{K}}\sum_{i\in A_s}\sum_{t=1}^{T}(\tilde{y}_{it} - \tilde{\mathbf{x}}_{it}'\hat{\boldsymbol{\beta}}_s(\lambda))^2$, the $\hat{\boldsymbol{\beta}}_s(\lambda)$'s are post Panel-CARDS and hierarchical clustering estimators, and here we make their dependence on λ 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 Ke et al. (2015), we use the normalized mutual information measure to assess the similarity between the estimated grouping structure $\hat{\mathcal{G}}$ and the true grouping structure $\mathcal{G}$. For two classifications/grouping structures $\mathcal{A} = \{A_1, A_2, \cdots\}$ and $\mathcal{B} = \{B_1, B_2, \cdots\}$ on the same set $\{1, 2, \ldots, N\}$, the normalized mutual information is defined as $\text{NMI}\quad(\mathcal{A},\mathcal{B}) = I(\mathcal{A},\mathcal{B})/\sqrt{H(\mathcal{A})H(\mathcal{B})}$, 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 $\text{NMI}\quad(\mathcal{A},\mathcal{B}) = 1$. In general, the more similar two classifications are, the closer their normalized mutual information value is to 1. We report $\text{NMI}\quad(\hat{\mathcal{G}},\mathcal{G})$ 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 Su et al. (2016) to define these criteria. The correct classification ratio is defined as $N^{-1}\sum_{k=1}^{K}\sum_{i\in\hat{G}_k}1\{\beta_i^0 = \alpha_{m(i)}^0\}$, where $m(i)$ denotes i's true group member. For the last three criteria, we focus on the estimates of the second slope coefficients. Let $\alpha_{\cdot 2}^0 \equiv (\alpha_{1,2}^0, \cdots, \alpha_{K^0,2}^0)'$ 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 $\alpha_{k,2}^0$s with weights N_k/N: $\sum_{k=1}^{K}\frac{N_k}{N}\text{RMSE}(\alpha_{k,2}^0)$. 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, $\tilde{\mathbf{B}} = (\tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2, \cdots, \tilde{\boldsymbol{\beta}}_N)$, we calculate the sample variance of each row of $\tilde{\mathbf{B}}$ and choose the two regressors with the largest variances for their coefficient estimates to construct the segmentations.

Figure 1 reports the classification results for DGP 1 for different combinations of N, T, and η. It shows the normalized mutual information between the estimated group structure $\hat{\mathcal{G}}$ and the true group structure $\mathcal{G}$ and suggests that, as T increases, the normalized mutual information between $\hat{\mathcal{G}}$ and $\mathcal{G}$ increases rapidly. When $T = 80$, the estimation is almost as good as the oracle for all values of η.. We also note that the performance of Panel-CARDS with $\eta = 2\%$ or 5% significantly improves that with $\eta = 0$, but a further increase of η does not necessarily lead to improved performance. Figure 2 reports the normalized mutual information for DGP 2 for various combinations of N, T, and η. The normalized mutual information patterns in Figure 2 are similar to those in Figure 1 for DGP 1. With respect to η, 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 (eight 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 of the Supporting Information.

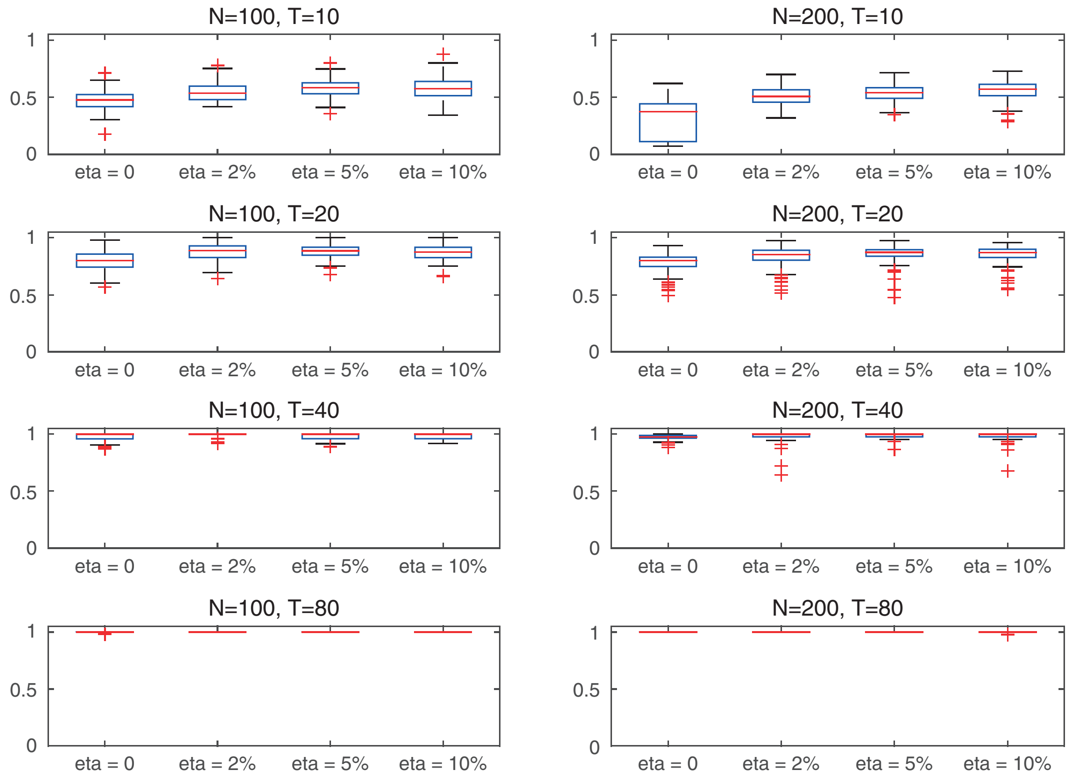

FIGURE 1 NMI of DGP 1 classification results using Panel-CARDS. The x-axis and y-axis mark the η and NMI values, respectively. The left and right columns report the results for $N = 100$ and $N = 200$, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

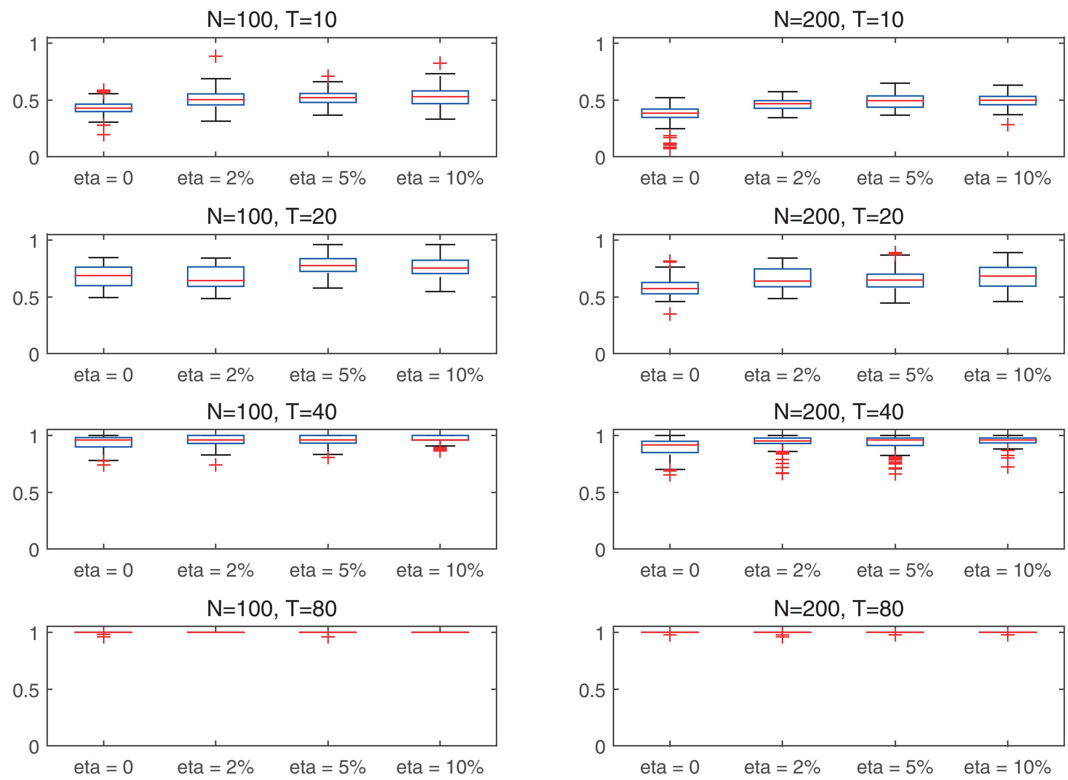

FIGURE 2 NMI of DGP 2 classification results using Panel-CARDS. (See Figure 1 for explanation) [Colour figure can be viewed at wileyonlinelibrary.com]

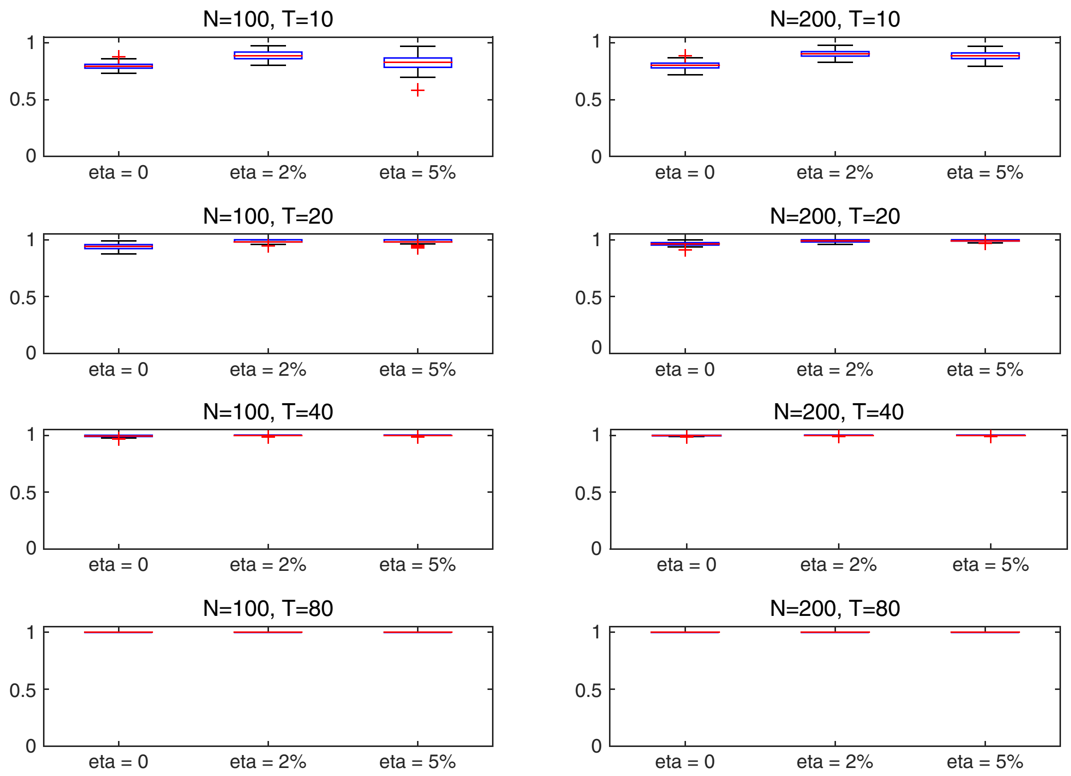

FIGURE 3 NMI of DGP 3 classification results using Panel-CARDS. (See Figure 1 for explanation) [Colour figure can be viewed at wileyonlinelibrary.com]

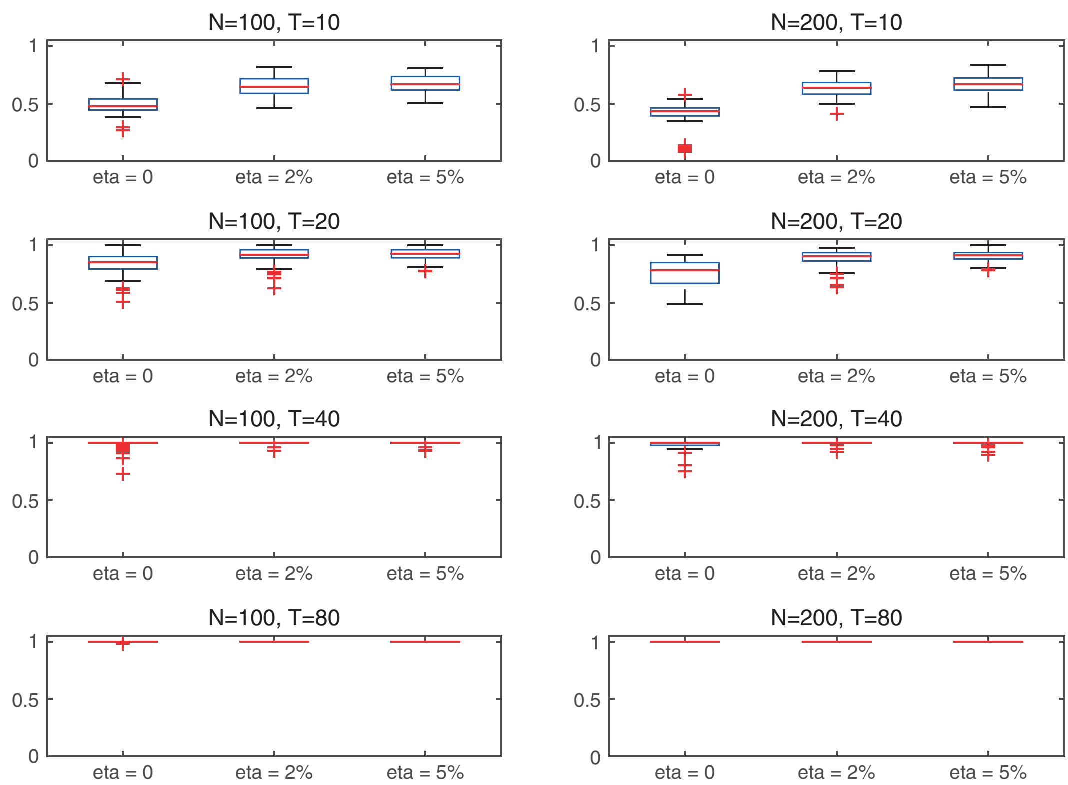

FIGURE 4 NMI of DGP 4 classification results using Panel-CARDS. (See Figure 1 for explanation) [Colour figure can be viewed at wileyonlinelibrary.com]

For the second slope coefficients $\{\alpha^0_{k,2}\}_{k=1}^{K}$ 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. 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,

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

			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

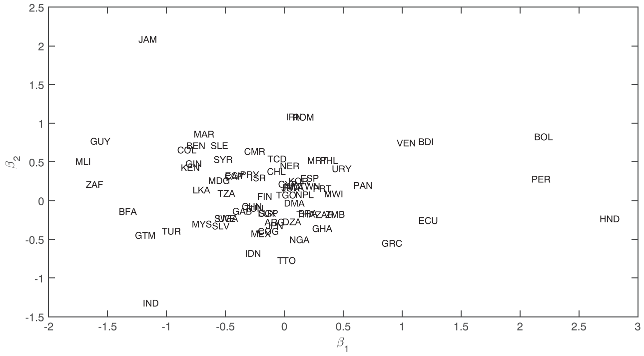

FIGURE 5 Scatter plot of preliminary estimates. See Table 10 in the Supporting Information for full names of these three-letter country codes

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

Following the lead of Acemoglu et al. (2008) and Bonhomme and Manresa (2015), we consider the following regression model with both individual and time fixed effects:

$$d_{it} = \beta_{i1} I_{i,t-1} + \beta_{i2} d_{i,t-1} + \mu_i + \gamma_t + \epsilon_{it}, i = 1, \cdots, N, t = 1, \cdots, T, \tag{16}$$

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, μ_i is the individual fixed effect, γ_t is the time fixed effect, ϵ_{it} is the error term, and β_{i1} and β_{i2} are the slope coefficients, which are assumed to be constant across countries in early studies. See Acemoglu et al. (2008) and Bonhomme and Manresa (2015) for detailed descriptions of the variables d_{it} and I_{it}.

We use the publicly available data that are used in Bonhomme and Manresa (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 5-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 Equation 16 by minimizing the nonpenalized objective function in Equation 14 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 of the Supporting Information 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 β_1 and positive β_2; Group 2 has negative β_1 and negative β_2; Group 3 has negative β_1 and positive β_2; Group 4 has positive β_1 and negative but insignificant β_2. These results are consistent with the scatter plot of the preliminary estimates in Figure 6 and suggest the effect of income on the level of democracy is not necessarily positive. Second, if

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

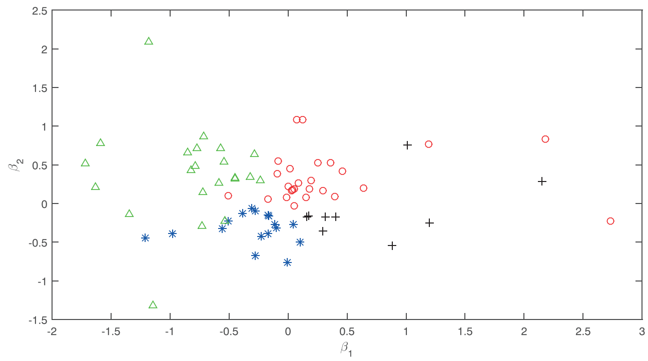

FIGURE 6 Scatter plot of classification results by using Panel-CARDS. As in Figure 5, the location of the scatter points indicate the value of preliminary estimates of β_1 and β_2. The circle, star, triangle, and plus symbols correspond to Group 1, Group 2, Group 3, and Group 4, respectively [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Regression results for groups 1–4 and the pooled one

	β_1			β_2			
	Estimates	**SE**	***t*-stat.**	**Estimates**	**SE**	***t*-stat.**	**LRE**
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 long run effect, which is defined as $\beta_1/(1 - \beta_2)$.

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 subclasses 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 Ke et al. (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 long-standing example of panel data research in economics—the international relationship between income and democracy. The methods identify four latent groups of countries that 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 No. MOE2012-T2-2-021 and the funding support provided by the Lee Kong Chian Fund for Excellence.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at [http://qed.econ.queensu.ca/jae/2018-v33.6/wang-phillips-su/].

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* (pp. 2101–2163). Amsterdam, Netherlands: Elsevier.

Browning, M., & Carro, J. (2007). Heterogeneity and microeconometrics modeling. In Blundell, R., Newey, W., & Persson, T. (Eds.), *Advances in economics and econometrics, theory and applications: Ninth world congress of the econometric society* (pp. 45–74). Cambridge, UK: Cambridge University Press.

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*. Berlin, Germany: 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.

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*). Cambridge, MA: National Bureau of Economic Research.

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.

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.

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

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

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 and Economic Statistics*. Advance online publication. https://doi.org/10.1080/07350015.2017.1340299

Sun, Y. (2005). Estimation and inference in panel structure models. (Working paper), Department of Economics, UCSD, San Diego, CA.

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.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Wang W, Phillips PCB, Su L. Homogeneity pursuit in panel data models: Theory and application. *J Appl Econ*. 2018;33:797–815. https://doi.org/10.1002/jae.2632