Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership Is Unknown

Chang-Ching Lin* Serena Ng†

Revised March 2010

Abstract

This paper proposes a new approach to estimate panel data models with group specific parameters when group membership is not known. We first create a set of "pseudo" threshold variables based on time series estimation of the individual specific parameters. We then use these variables to stratify individuals. The problem of parameter heterogeneity is turned into estimation of a panel threshold model in which the threshold variables are themselves being estimated. We show that individuals can be consistently sorted into groups distinguished by parameter heterogeneity when N and T are large. Results are compared to the K-means algorithm adapted to panel data regressions with fixed effects.

Keywords: Parameter Heterogeneity, Threshold Models, Cluster Analysis.

*Institute of Economics, Academia Sinica, 128 Academia Road, Sec 2, Taipei 115, Taiwan. Tel: 886-2-27822791 ext. 301. E-mail: lincc@econ.sinica.edu.tw

†Department of Economics, Columbia University, 420 West 118 St., New York, NY 10027 USA tel: 212-854-5488 Email: serena.ng@columbia.edu.

We thank three anonymous referees and the editor for many helpful suggestions and comments. The second author acknowledges financial support from the National Science Foundation (SES-0549978).

1 Introduction

This paper considers estimation of panel data models when the slope parameters are heterogeneous across groups, but that group membership is not known to the econometrician. We propose a data dependent method that groups individuals according to consistent estimates of the slope coefficients at the individual level. Our analysis proceeds in three steps. First, we use time series estimates of the individual slope coefficients to form a set of "pseudo" threshold variables. Second, the threshold value is estimated and used to partition the sample into groups. Third, the model is re-estimated by pooling observations within groups. Thus, units within a group have homogeneous parameters but the parameters are heterogeneous across groups. We turn the problem of identifying group membership into one of estimating a threshold panel regression, where the threshold variable is itself being estimated. We refer to this as a 'pseudo' threshold approach.

Panel data models often take parameter homogeneity as a maintained assumption even though evidence against it is not difficult to find. Using data on US manufacturing, Burnside (1996) rejects homogeneity of the parameters in the production function. Lee, Pesaran, and Smith (1997) reject the hypotheses that the rate of technological growth and the rate of convergence of per capita output to the steady state level are the same across countries. Hsiao and Tahmiscioglu (1997) find heterogeneity in the parameters of equations that describe investment dynamics and observed that such differences cannot be explained by commonly considered firm characteristics. Barsky, Juster, Kimball, and Shapiro (1997) find substantial heterogeneity in the rate of time preference (say, ρ) and the elasticity of intertemporal substitution (say, τ) among respondents of the Health and Retirement Survey. Thus, if r is the real interest rate and c is consumption, and theory implies $\Delta \log c = \tau(r - \rho)$ for a particular household, the parameters α_0 and α_1 in a panel regression model $\Delta \log c_{it} = \alpha_0 + \alpha_1 r_t$ should vary across individuals. Lawrance (1991) allows the discount rate and the rate of time preference in the consumption Euler equation to differ between rich and poor households and along demographic lines. Guvenen (2009) finds that allowing stockholders to have a higher elasticity of substitution than non-stockholders can explain a number of asset pricing phenomena. Carroll and Samwick (1997) find that labor income dynamics are heterogeneous across education groups.

As Browning and Carro (2007) point out, there is usually much more heterogeneity than empirical researchers allow. Robertson and Symons (1992) use monte carlo experiments to show that

the bias in the Anderson and Hsiao (1982) estimator can be severe when the parameters vary across individuals but this variation is not allowed for in estimation. While fixed effects estimation handles heterogeneity in the intercept, few methods are available to allow for heterogeneity in the slope parameters. The issue is that assuming complete parameter heterogeneity would reduce the problem to time series estimation on a unit by unit basis which does not take advantage of the panel structure of the data. Partitioning the data into groups is an immediate approach that permits some pooling and yet still accommodates heterogeneity in the regression function.

If we know which group each unit belongs, we can simply do split sample linear regressions. The main obstacle is that group membership is not always known. One approach is to use a priori and observed information to organize units into groups, but the approach is not objective. For example, should one use income or wealth to classify who is rich and who is poor, and what is the cut-off point? Furthermore, units differ in many dimensions and there are often several ways to partition the data.

Instead of using a priori information, we let the data determine the grouping. Our proposed 'top down' method provides a simple characterization of how the units respond differently to the covariates. In the Euler equation example above, our analysis would sort households into a group with high and a group with low intertemporal elasticity of substitution. This is to be distinguished from a 'bottom up' approach that forms groups by explicitly specifying the sources of parameter heterogeneity.

In addition to the pseudo-threshold method, we also adapt the K-means algorithm to panel regressions. The K-means is a popular clustering algorithm that shuffles observations into appropriate groups until the within cluster variances is minimized. The method is usually used to cluster a set of data points and we are unaware of its application to regression analysis. The main difference between the K-means and our method is that we use information about the individual slope coefficients to do the shuffling, which is less of a black box, and is computationally less intensive.

The remainder of this paper is organized as follows. After a review of related work, Section 3 presents the pseudo threshold method. Extension to the case of multiple groups and multiple covariates is considered in Section 4. Adaptation of the K-means method to panel regressions is given in Section 5. Simulations are then presented. As an application, we apply the methods to study economic growth across countries.

2 Related Literature and the Econometric Framework

There are many methods that allow for heterogeneity in the regression function. For example, classification analysis in the form of regression trees (CART) involves repeated subdivisions of a group of observations on the basis of optimal cut-off points of the covariates. Parameters can differ across nodes if desired. Durlauf and Johnson (1995) use CART to understand why some countries have high growth while others have low growth. Regression splines also allow for group specific parameters; it does so using a priori information to form knot points. For example, households are considered liquidity constrained if their wealth exceeds a certain level, while firms are grouped by their capital intensity. In cross-country analysis, groups are sometimes formed depending on whether a country is a member of the OECD. Spatial information has also been used for sample splitting. However, in regressions with multiple covariates as is often the case, there are often several logical but not mutually exclusive ways to partition the sample. Furthermore, groups that are deemed to be economically meaningful need not be optimal from a statistical point of view.

An alternative method of introducing flexibility to the regression function is to allow the coefficients to be heterogeneous by parameterizing them as a function observed characteristics as in Alvarez, Browning, and Ejrnæs (2006). The analysis would necessarily depend on the parametric functions used. Alternatively, a random coefficient model[1] can provide efficient estimates for the average effect of the covariates on the endogenous variable, but is uninformative about the response at a more disaggregated level, which is sometimes an object of interest. Indeed, it is not useful to talk about the individual parameters of the random coefficient model in a frequentist setting since they are treated as random variables. Maddala, Trost, Li, and Joutz (1997) discuss a Bayesian method that shrinks the individual estimates toward the estimator of the overall mean.

Allowing the parameters to be homogeneous within groups but heterogeneous across groups is a form of model based clustering. Clustering analysis partitions a set of data, $x_i, i = 1, \ldots N$, into G groups according to how near they are to one another.[2] This is to be distinguished from classification analysis in which the objective is to understand how the predefined groups differ. Allowing the parameters to be different across groups is also different from allowing the marginal effects to differ through splitting the sample on the basis of values of the regressors.

The simplest way to form clusters is to plot the unconditional mean of the data $\hat{\beta}_i = \bar{y}_i$ and

[1] See, for example, Swamy (1970) and Hsiao and Pesaran (2004).

[2] See, for example, Fraley and Raftery (2002), Hall, Muller, and Wang (2006), and Chiou and Li (2007).

then 'eyeball' to see when $\hat{\beta}_i$ abruptly shifts from one mean to another.[3] Such a graphical approach is often a useful diagnostic, but does not permit formal statistical statements to be made. Model based clustering takes as starting point that a set of data with a group structure is generated by a mixture of distributions such that an observation drawn from sub-population g has density $f_g(x_i|\beta_g)$. If q_i is the identifying label, i.e. $q_i = g$ if unit i belongs to group g, then one can maximize the likelihood $L(x;\theta,q) = \prod_{i=1}^{N}\prod_{t=1}^{T} f_{q_i}(x_{it};\beta)$ with respect to β using the EM algorithm. The unknown identifier q_i would be estimated by the empirical probability of the group to which unit i belongs. The method can be cumbersome if N is large because we need to consider up to 2^N possible combinations. Sun (2005) modifies the EM algorithm to restrict the units in a cluster to share common parameters. A logit regression is used to infer to which group unit i belongs, and weighted least squares method is used to estimate the group parameters. Consistency and asymptotic normality of the maximum likelihood estimator are proved under the assumption that N is large and T is fixed. Juárez and Steel (2010) propose a Bayesian method assuming that the errors are cross-sectionally homogeneous, and that are independently drawn from a t-distribution, and the individual-specific fixed effects are normally distributed. Our approach does not impose parametric assumptions on the functional forms or on the errors.

Our approach is a form of model based clustering, but our primary objective is not identifying the clusters per se. Rather, we want to pool 'similar' observations to estimate the parameters of the model, where similarity is defined in terms of the slope coefficients. We consider a balanced panel of data with observations on N cross-section units over T time periods. There are K regressors and G clusters, and to introduce the main idea, we start with the simple case of $K = 1$ and $G = 2$. Let $I^0 = (I_1^0, I_2^0)$ be indicator variables that denote true group membership and let N_1^0 and N_2^0 denote the number of individuals in clusters I_1^0 and I_2^0, respectively. The data are assumed to be well approximated by the model:

$$\begin{aligned} \widetilde{y}_{it} &= \alpha_i + \widetilde{x}_{it}\beta_i + \widetilde{e}_{it}, \quad \widetilde{e}_{it} \sim (0, \sigma_i^2) \\ &= \alpha_i + \widetilde{x}_{it}B_1 1(i \in I_1^0) + \widetilde{x}_{it}B_2 1(i \in I_2^0) + \widetilde{e}_{it}, \end{aligned} \tag{1}$$

where α_i is the individual fixed effects, and $\widetilde{x}_{it}$ is a vector of predetermined variables. The coefficients for the two clusters are given by B_1 and B_2. That is, $\beta_i = B_1$ if $i \in I_1^0$ and $\beta_i = B_2$ if $i \in I_2^0$. Without loss of generality, we assume that $B_1 < B_2$. The case of homogeneous parameters is

[3] See, for example, Henderson and Russell (2005).

obtained if $B_1 = B_2$, and the case of complete heterogeneity is obtained when there are N clusters each consisting of only one unit.

We control for the fixed effects by demeaning. Let $z_{it} = \widetilde{z}_{it} - \overline{\widetilde{z}}_i$ where $\overline{\widetilde{z}}_i = \frac{1}{T}\sum_{t=1}^{T} \widetilde{z}_{it}$, and $\widetilde{z}_{it}$ can be $\widetilde{y}_{it}, \widetilde{x}_{it}$, or $\widetilde{e}_{it}$. We can rewrite the model as

$$y_{it} = x_{it}B_1 1(i \in I_1^0) + x_{it}B_2 1(i \in I_2^0) + e_{it}. \tag{2}$$

We will be concerned with both individual and pooled estimates of β_i. For each i, let $\hat{\beta}_i$ be the least squares estimate of β_i obtained from a time series regression of y_{it} on x_{it}. A pooled estimate of β_i is obtained by considering the model

$$\widetilde{y}_{it} = \alpha_i + \widetilde{x}_{it}B_\omega + \widetilde{e}_{it}. \tag{3}$$

The pooled fixed effects estimator is then defined as

$$\hat{B}_\omega = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}' y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}' x_{it}} = \sum_{i=1}^{N}\left(\frac{\sum_{t=1}^{T} x_{it}' x_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}' x_{it}}\right)\hat{\beta}_i.$$

Let $I = (I_1, I_2)$ denote a group membership other than $I^0 = (I_1^0, I_2^0)$. For $j, k = 1, 2$, let N_{kj} be the number of individuals assigned to group j when they truly belong to group k. Let $N_1 = N_{11} + N_{21}$, $N_2 = N_{22} + N_{12}$ and let $N_s = N_s(I, I^0) = N_{12} + N_{21}$ be the number of misclassified units. Also, let $\widetilde{x}_i = (\widetilde{x}_{i1}', \ldots, \widetilde{x}_{iT}')'$. The following assumptions will be used throughout for a panel data model with strictly exogenous regressors.

Assumption 1: For all i's and t's, (a) $\widetilde{e}_{it}|\widetilde{x}_i \sim IID(0, \sigma_i^2)$. $\{\widetilde{e}_{it}\}$ are cross-sectionally independently distributed and uncorrelated with B_1 and B_2. (b) $\max_{1\le i\le N}\sigma_i^2$ is finite and $\min_{1\le i\le N}\sigma_i^2 > 0$. Furthermore, (y_{it}, x_{it}) are jointly stationary.

Assumption 2: For $j = 1, 2$, $N_j^0/N > 0$ and $N_1^0/N \to \pi$ with $0 < \pi < 1$.

Assumption 3: (a) For each i, $\hat{Q}_i = T^{-1}\sum_{t=1}^{T} x_{it}' x_{it}$ is finite and positive definite and $\max_{1\le i\le N} \mathrm{E}||\hat{Q}_i||$ is finite with $\hat{Q}_i \xrightarrow{p} Q_i > 0$ as $T \to \infty$, where Q_i is a non-stochastic positive definite and $\max_{1\le i\le N} \mathrm{E}||Q_i||$ is finite. (b) Let I_* be a nonempty subset of the whole sample and let N_* denote the number of units in I_*. Assume that $N_*^{-1}\sum_{i\in I_*}\hat{Q}_i$ has the minimal eigenvalue bounded away from zero in probability as $(N_*, T) \to \infty$ jointly.

Assumption 1 is commonly imposed in panel data models with fixed effects. Assumption 2 ensures that the groups are not degenerate. Assumption 3 is an identification condition. Cross-section dependence can be entertained if we allow $\widetilde{e}_{it}$ to have a factor structure. This will change the

formulation of the individual and the pooled regressions but our proposed method remains valid. To focus on the issue of parameter heterogeneity, cross-section independence is the maintained assumption. Cross-section dependence will be considered in the application.

A dynamic panel model obtains when $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$. To accommodate these models, the following assumptions will be necessary:

Assumption D1: For all i's and t's, (a) $\widetilde{e}_{it} \sim (0, \sigma_i^2)$ are cross-sectionally and serially independently distributed, independent of the initial values y_{i0}, with finite moments up to fourth order, and are uncorrelated with B_1 and B_2. (b) $\max_{1\leq i\leq N} \sigma_i^2$ is finite and $\min_{1\leq i\leq N} \sigma_i^2 > 0$. Furthermore, $|B_1| < 1$, $|B_2| < 1$, and $N^{-1}\sum_{i=1}^{N} \alpha_i = O(1)$.

Assumption D2: (a) For $j = 1, 2$, $N_j^0/N > 0$ and $N_1^0/N \to \pi$ with $0 < \pi < 1$. (b) $0 \leq \lim N/T < \infty$ as N and T diverge jointly.

Assumption D3: For all i's and t's, the initial observations satisfy $\widetilde{y}_{i0} = \alpha_i/(1-\beta_i) + \widetilde{u}_{i0}$, where $\widetilde{u}_{i0} \sim (0, \sigma_{v,i}^2)$ are cross-sectionally independently distributed, independent of $\widetilde{e}_{it}$, uncorrelated with B_1 and B_2, with $0 < \sigma_{v,i}^2 < \infty$, and with finite moments up to fourth order.

Unlike the panel data model with strictly exogenous regressors, Assumption D2(b) ensures the asymptotic bias of the fixed effects estimator in a dynamic panel remains bounded as N and T diverge jointly. See Hahn and Kuersteiner (2002), Alvarez and Arellano (2003), and Pesaran and Yamagata (2008).

Lemma 1 *Let $B_\omega = \omega B_1 + (1-\omega)B_2$, where $\omega = (\sum_{i=1}^{N} \hat{Q}_i)^{-1} \sum_{i\in I_1^0} \hat{Q}_i$. Suppose that Assumptions 1–3 hold or that $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$ and Assumptions D1–D3 hold. Then (a) $\sqrt{NT}(\hat{B}_\omega - B_\omega) = O_p(1)$. (b) Let $\omega_0 = \text{plim}_{N\to\infty}\omega$. $\hat{B}_\omega \xrightarrow{p} \omega_0 B_1 + (1-\omega_0)B_2$.*

Lemma 1 shows that $\hat{B}_\omega$ is consistent for the population mean B_ω even though the regression model (3) is misspecified when the true model has heterogeneous slope parameters.

3 A Threshold Approach

Goldfeld and Quandt (1973) were the first to use threshold variables, also referred to as transition variables, to form clusters. They considered a model in which the clusters are determined by a linear function of several transition variables. They proposed a D-method within the maximum likelihood framework to enable estimation of the parameters in the transition function. The D

method assumes deterministic switching of regimes, and stands in contrast to the λ-method in which units are assigned to regimes in a random manner. A more popular idea, also due to Goldfeld and Quandt (1973), is to partition a data set based on a known threshold variable taking on an unknown threshold value. Threshold autoregressive models and structure break models are variations of this idea in a time series context.

Suppose we can find a variable q_i that, along with a set of cut-off parameter values Γ^0, will lead to perfect information about I_1^0 and I_2^0 in the sense that $i \in I_1^0$ if $q_i \leq \gamma^0$ for any $\gamma^0 \in \Gamma^0$ and $i \in I_2^0$ otherwise. Then (1) can be written as

$$\widetilde{y}_{it} = \alpha_i + \widetilde{x}_{it} B_1 1(q_i \leq \gamma^0) + \widetilde{x}_{it} B_2 1(q_i > \gamma^0) + \widetilde{e}_{it}. \tag{4}$$

Hansen (1999) considered threshold panel regressions where the sample is split according to whether q_{it} is less than some γ. In his analysis, q_{it} is an observed variable that is often one of the $\widetilde{x}_{it}$, and it is time varying. Unit i can be in one group in period t if $q_{it} \geq \gamma^0$, but is in another group in period $t+1$ if $q_{it+1} < \gamma^0$. In contrast, our threshold variable q_i is not observed, and group structure does not change over time. Because of these differences, we call q_i a 'pseudo threshold variable' and γ the 'pseudo threshold parameter' to distinguish them from the usual definitions used in the threshold literature.

If q_i and Γ^0 are both known, estimates of B_1 and B_2 can be obtained using a threshold, or split-sample, regression. Observations with $q_i \leq \gamma^0$ for any $\gamma^0 \in \Gamma^0$ can be pooled to estimate B_1, while observations with $q_i > \gamma^0$ can be pooled to estimate B_2. The problem, however, is that neither q_i nor Γ_0 is observed. We first discuss how to estimate γ assuming q_i is known. We then consider two possible choices of q_i.

3.1 Estimation of γ when q_i is Known

When q_i is known and exogenous but Γ^0 is not observed, a $\gamma^0 \in \Gamma^0$ can be estimated as follows. Order the observations by q_i. For a given $\gamma \in \Gamma = [q_{\min}, q_{\max}]$, let $\hat{B}_1(\gamma)$ and $\hat{B}_2(\gamma)$ denote the least squares estimator of B_1 and B_2 using observations with $q_i \leq \gamma$ and $q_i > \gamma$ respectively. Then

$$\widetilde{\gamma} = \arg \min_{\gamma \in [q_{\min}, q_{\max}]} S_{NT}(\gamma), \tag{5}$$

where the sum of squared residuals is defined as

$$\begin{aligned} S_{NT}(\gamma) &= \sum_{i=1}^{N}\sum_{t=1}^{T}\Big(y_{it} - x_{it}\hat{B}_1(\gamma)1(q_i \le \gamma) - x_{it}\hat{B}_2(\gamma)1(q_i > \gamma)\Big)^2 \\ &= \sum_{i|q_i\le\gamma}\sum_{t=1}^{T}(y_{it} - x_{it}\hat{B}_1(\gamma))^2 + \sum_{i|q_i>\gamma}\sum_{t=1}^{T}(y_{it} - x_{it}\hat{B}_2(\gamma))^2. \end{aligned}$$

Since q_i can be used to order the data, this means that if q_i is less than some trial value of γ and the i-th unit is classified into group 1, any unit j with $q_j < q_i$ will also be classified in the group. Using q_i to split the unordered sample at some γ is isomorphic to splitting the ordered sample at some observation i^* that has $q_{i^*} = \gamma$. Therefore, even though there are 2^N possible groupings of the data, we only need to consider at most $N-1$ possible values of γ.

Let $N_{kj}(\gamma)$ be the number of units that belong to group k but are classified into group j when the sample is partitioned at γ. Notice that when too many units are put in Group 1 (and thus $N_{21}(\gamma) > 0$), then it will also be the case that $N_{12}(\gamma) = 0$. Thus, one of the misclassified set is always empty. A unit misclassified into Group 1 will contribute a larger squared error than if the unit was put into Group 2 since B_2 is closer to $\hat{B}_2(\gamma)$ than $\hat{B}_1(\gamma)$. Minimizing $S_{NT}(\gamma)$ should then yield a $\widetilde{\gamma}$ that also minimizes the number of misclassified units.

Theorem 1 *Suppose the data are generated by (1) and q_i is exogenous and observed. Suppose that $\widetilde{\gamma}$ is obtained from (5). Then for $j = 1, 2$, $\hat{B}_j(\widetilde{\gamma}) - B_j = o_p(1)$. If $B_2 - B_1 = \nu T^{-\alpha}$ with $0 < ||\nu|| < \infty$ and $0 \le \alpha < 1/2$, then $N_s(\widetilde{\gamma}) = O_p(T^{-1+2\alpha})$.*

If the trial value of γ is too low, $\hat{B}_2(\gamma)$ will be calculated with some observations from group 1 and will not be consistent for B_2. Similarly, at too high a value of γ, $\hat{B}_1(\gamma)$ will be calculated with observations from group 2, and hence will not be consistent for B_1. We only need to consider where to position γ in relation to the N ordered observations of q_i, denoted $q_{[i]}$. There will be a $\widetilde{\gamma}$ that minimizes the size of the misclassified set. In fact, any $\widetilde{\gamma} \in [q_{[i^*]}, q_{[i^*]+1})$, where i^* is such that $\widetilde{q}_{[i^*]} = \widetilde{\gamma}$, will yield the same clusters. For fixed $B_2 - B_1 = O(1)$, Theorem 1 implies that the maximum misclassification rate is $N_s(\widetilde{\gamma})/N = O_p(\frac{1}{NT})$. If $B_2 - B_1$ is in the $T^{-\alpha}$ neighborhood of zero, the misclassification rate is $N_s(\widetilde{\gamma})/N = O_p(\frac{1}{NT^{1-2\alpha}})$. Thus the misclassification rate tends to zero as $N, T \to \infty$ jointly.

Given $\widetilde{\gamma}$, the two groups can be estimated as $I_1(\widetilde{\gamma}) = \{i|q_i \le \widetilde{\gamma}\}$ and $I_2(\widetilde{\gamma}) = \{i|q_i > \widetilde{\gamma}\}$. Once group membership is consistently estimated, units within a group can be pooled to yield more

efficient cluster-specific parameters. Consistency and asymptotic normality of $\hat{B}_1$ and $\hat{B}_2$ can be established by treating $\widetilde{\gamma}$ as though it was known. While likelihood based cluster analysis yields a probability that unit i belongs to a group, group membership is known once we can find an appropriate q_i.

3.2 A Two-step Pseudo Threshold Approach

In practice, q_i is not observed. We propose to replace q_i by some $\hat{q}_i$ that has the same information as q_i in the sense that $\hat{q}_i \leq \gamma$ when $q_i \leq \gamma$ as $T \to \infty$. To simplify notation, hereafter, all variables indexed by i are assumed to be ordered once q_i is estimated. Given $\hat{q}_i$, the problem is to find an estimate of γ. Let

$$\hat{\gamma} = \arg \min_{\gamma \in [\hat{q}_{\min}, \hat{q}_{\max}]} S_{NT}(\gamma, \hat{q}), \tag{6}$$

where

$$S_{NT}(\gamma, \hat{q}) \quad = \quad \sum_{i|\hat{q}_i \leq \gamma} \sum_{t=1}^{T} (y_{it} - x_{it}\hat{B}_1(\gamma))^2 + \sum_{i|\hat{q}_i > \gamma} \sum_{t=1}^{T} (y_{it} - x_{it}\hat{B}_2(\gamma))^2. \tag{7}$$

The two groups are then estimated as

$$\hat{I}_1 = \{i|\hat{q}_i \leq \hat{\gamma}\} \quad \text{and} \quad \hat{I}_2 = \{i|\hat{q}_i > \hat{\gamma}\}.$$

To motivate our choices of $\hat{q}_i$, consider letting $q_i = \beta_i - B_\omega$. It is easy to see that

$$q_i = \beta_i - B_\omega = \begin{cases} \beta_i - B_1 - (1-\omega)(B_2 - B_1) & \text{for } i \in I_1^0, \\ \\ \beta_i - B_2 + \omega(B_2 - B_1) & \text{for } i \in I_2^0, \end{cases}$$

where $\omega = (\sum_{i=1}^{N} \hat{Q}_i)^{-1} \sum_{i \in I_1} \hat{Q}_i$. Now $\beta_i - B_1 = 0$ if $i \in I_1^0$ and $B_2 \neq B_1$ by assumption. Thus, $q_i = -(1-\omega)(B_2 - B_1) < 0$ if $i \in I_1^0$. On the other hand, $q_i = \omega(B_2 - B_1) > 0$ if $i \in I_2^0$. The pseudo variable $q_i = \beta_i - B_\omega$ along with any $\gamma^0 \in [-(1-\omega)(B_2 - B_1), \omega(B_2 - B_1))$ completely summarizes group membership. The procedure can be further simplified by noting that B_ω is common across i. This implies that $q_i = \beta_i$ along with any $\gamma^0 \in \Gamma^0 = [B_1, B_2)$ will also identify group membership. As noted by one of the referees, a convenient choice of γ^0 in this case is B_ω.

Although β_i is not known, it can be consistently estimated using the time series observations on unit i only. Furthermore, $\hat{B}_\omega$ obtained from a pooled regression is also consistent for B_ω. It follows that $\hat{q}_i = q_i + O_p(T^{-1/2})$ for (i) PSEUDO1: $\hat{q}_i = \hat{\beta}_i - \hat{B}_\omega$, $q_i = \beta_i - B_\omega$ and (ii) PSEUDO2:

$\hat{q}_i = \hat{\beta}_i$, $q_i = \beta_i$.[4] The main difference between the two is that under PSEUDO1, γ is estimated from minimization of $S_{NT}(\gamma, \hat{q})$. Under PSEUDO2, γ is estimated by the pooled estimate of B, *ie.* $\hat{B}_\omega$.

In Theorem 1, we have shown that when q_i is known and exogenous the classification error rate is $O_p(N^{-1}T^{-1+2\alpha})$, and if $B_2 - B_1$ is fixed, $P(N_s(\widetilde{\gamma})/N|q_i) = O_p((NT)^{-1})$. Although $\hat{\beta}_i$ and $\hat{B}_\omega$ are both subject to sampling variability, the classification error rate of our pseudo threshold method still converges to zero, albeit at a slower rate. Consider first PSEUDO2 with $\hat{\gamma} = \hat{B}_\omega$. Under Assumptions 1–3, [5] we have

$$\begin{aligned} P(\hat{\beta}_i > \hat{B}_\omega|\beta_i = B_1) &= P\bigg(\sqrt{T}(\hat{\beta}_i - B_1) > \sqrt{T}(\hat{B}_\omega - B_1)\bigg|\beta_i = B_1\bigg) \\ &= P\bigg(\sqrt{T}(\hat{\beta}_i - B_1) > \sqrt{T}\bigg[(1-\omega)(B_2 - B_1) + O_p(\frac{1}{\sqrt{NT}})\bigg]\bigg|\beta_i = B_1\bigg) \\ &= P\bigg((\hat{\beta}_i - B_1) > \bigg[(1-\omega)\nu T^{-\alpha+1/2} + O_p(\frac{1}{\sqrt{N}})\bigg]\bigg|\beta_i = B_1\bigg). \end{aligned}$$

Similarly,

$$P(\hat{\beta}_i < \hat{B}_\omega|\beta_i = B_2) = P\bigg((\hat{\beta}_i - B_2) > \bigg[\omega\nu T^{-\alpha+1/2} + O_p(\frac{1}{\sqrt{N}})\bigg]\bigg|\beta_i = B_2\bigg).$$

Let $N_s(\hat{\gamma})$ be the number of misclassified units given $\hat{\gamma}$ and $\hat{q}_i$. Now $N_s(\hat{\gamma}) = \sum_{i\in I_1^0} 1(\hat{\beta}_i > \hat{B}_\omega) + \sum_{i\in I_2^0} 1(\hat{\beta}_i < \hat{B}_\omega)$. Thus, $E(N_s(\hat{\gamma})/N) = P(\hat{\beta}_i > \hat{B}_\omega|\beta_i = B_1) + P(\hat{\beta}_i < \hat{B}_\omega|\beta_i = B_2) = O(T^{-1+2\alpha})$.

For PSEUDO1 with $\hat{\gamma}$ estimated from a threshold regression, we have

$$P(N_s(\hat{\gamma})/N) = O_p(\max(N^{-1}T^{-1+2\alpha}, T^{-1+2\alpha})) = O_p(T^{-1+2\alpha}).$$

Here, the rate of $N^{-1}T^{-1+2\alpha}$ arises from having to estimate γ, while the rate of $T^{-1+2\alpha}$ arises from having to estimate q_i. The overall correct classification rate is then dominated by how precisely we can estimate q_i.

Theorem 2 *Suppose the data are generated by (1). Let q_i be estimated by PSEUDO1 ($\hat{q}_i = \hat{\beta}_i - \hat{B}_\omega$) or PSEUDO2 ($\hat{q}_i = \hat{\beta}_i$). For $B_2 - B_1 = \nu T^{-\alpha}$ with $0 < ||\nu|| < \infty$ and $0 \le \alpha < 1/2$, $N_s(\hat{\gamma})/N \to 0$ as $(N,T) \to \infty$ jointly.*

[4] We also consider using $\hat{q}_i = \frac{(\hat{\beta}_i - \hat{B}_\omega)}{\hat{\sigma}_i \hat{Q}_i^{-1/2}}$ where $\hat{Q}_i$ is defined in Assumption 3(a) with $\hat{Q}_i \to Q_i > 0$ as $T \to \infty$, $\hat{\sigma}_i^2 = \frac{1}{T-G-1}\sum_{t=1}^T \hat{e}_{it}^2$, and $\hat{e}_{it} = y_{it} - x_{it}\hat{\beta}_i$. By standardizing the deviation between the individual estimate of β_i and an estimate of B_ω, we account for the sampling variability arising from time series estimation of β_i as well as fixed effects estimation of B_ω. This threshold variable gives more precise classification when there is substantial heterogeneity in σ_i.

[5] Similar results can be obtained under Assumptions D1–D3 when $x_{it} = y_{i,t-1}$. See the proof to Theorem 2.

In this framework, $\hat{q}_i - q_i$ only has a convergence rate of $O_p(T^{-1/2})$. A consequence of the two step procedure is that when T is small, the classification error can be high.

3.3 Extension to Multiple Regressors

We now turn to the case when there are $K > 1$ regressors. Note first that if a subset of the K parameters are suspicious of being different across groups, a case which we refer to as partial parameter homogeneity, the analysis in the previous section is still valid. For example, if the second slope coefficient varies between groups, we can let $\hat{q}_i = \hat{\beta}_{i2}$.

More difficult to handle is the case of complete parameter heterogeneity which arises when all K coefficients are group specific. To see why this is more involved, suppose there are two regressors, $x_{1,it}$ and $x_{2,it}$ and there are $G = 2$ clusters. Let $B_1 = (B_{11}, B_{12})'$ and $B_2 = (B_{21}, B_{22})'$ be the slope parameters for Group 1 and Group 2, respectively. Suppose first that for $j, k, = 1, 2$, $j \neq k$, we have $B_{j1} > B_{k1}$ and $B_{j2} > B_{k2}$. Since both parameters are strictly larger in one group than in another group, a natural pseudo transition variable is $\hat{\beta}_i^+ = \hat{\beta}_{i1} + \hat{\beta}_{i2}$. But this pseudo threshold variable does not always work! For example if $(B_{11}, B_{12}) = (0.8, 1)$ and $(B_{11}, B_{12}) = (1, 0.8)$, we have $B_{11} + B_{12} = B_{21} + B_{22}$. Thus when $B_{j1} > B_{k1}$ but $B_{j2} < B_{k2}$, the sum of the coefficients is no longer a sufficient statistic for group membership. In this case we need to consider the transition variable $\hat{\beta}_i^- = \hat{\beta}_{i1} - \hat{\beta}_{i2}$. Although we can expect $\hat{\beta}_i^+$ and $\hat{\beta}_i^-$ to separate those $i \in I_1^0$ from those $i \in I_2^0$ when T is large, we first need to determine the sign of the coefficients and then find a way uses this information to classify units. In an earlier version of this paper, we used the Goodman-Kruskal's gamma statistic to measure the association between pairs of concordant (same sign) and discordant data (opposite sign) data. As this sample statistic itself has variability, the procedure remains heuristic even though it works reasonably well in simulations.

A simpler and more effective approach is to recognize that even in the case of complete parameter heterogeneity, we can still split the sample using $\hat{\beta}_{ik}$ for some $k = 1, \ldots, K$ since each component of $\hat{\beta}_i = (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{iK})$ is informative about group membership. The only issue that remains is which particular component of $\hat{\beta}_i$ to use. We let the data speak by considering each component as a possible candidate and choose the one that minimizes the sum of squared residuals. More precisely, for each $i = 1, \ldots, N$, $k = 1, \ldots, K$, let $\hat{\gamma}_k$ be estimated from (6) for PSEUDO1 with

$\hat{q}_{ik} = \hat{\beta}_{ik} - \hat{B}_{\omega k}$ or $\hat{\gamma}_k = \hat{B}_{\omega k}$ for PSEUDO2 with $\hat{q}_{ik} = \hat{\beta}_{ik}$. Compute

$$S_{NT,k}(\hat{\gamma}_k) \quad = \quad \sum_{i|\hat{q}_{ik} \leq \hat{\gamma}_k} \sum_{t=1}^{T} (y_{it} - x_{it}' \hat{B}_1(\hat{\gamma}_k))^2 + \sum_{i|\hat{q}_{ik} > \hat{\gamma}_k} \sum_{t=1}^{T} (y_{it} - x_{it}' \hat{B}_2(\hat{\gamma}_k))^2,$$

where $x_{it}, \hat{B}_1, \hat{B}_2$ are $K \times 1$ vectors. The best threshold variable is $\hat{\beta}_{ik^*}$ where

$$k^* \quad = \quad \min_k S_{NT,k}(\hat{\gamma}_k). \tag{8}$$

The appeal of this approach is generality, since the procedure is the same for any K, and it works for partial or complete parameter heterogeneity.

4 K-means Clustering

Suppose that we observe y_i, $i = 1, \ldots, N$ and there are no covariates. The K-means algorithm produces G clusters by moving unit i to an appropriate group to minimize the sum of squared deviations between the units and the centroids.[6] The K-means method can be sensitive to the initial choice of the centroids and is not guaranteed to find the global minimizer. In spite of these shortcomings, the algorithm is quite popular in applied statistical work, though we are unaware of its application to a regression setting. We now modify the K-means method to allow for covariates.

Suppose that there are two groups and consider the transformed fixed effects model in (2). The algorithm consists of repeating the following steps.

1. Randomly assign individuals into two groups $\{\ddot{I}_1, \ddot{I}_2\}$, and calculate fixed effects estimator $(\ddot{B}_1, \ddot{B}_2)$ based on $\{\ddot{I}_1, \ddot{I}_2\}$

2. Repeat (a) and (b) until no individual is changed from one group to another: (a) Calculate $SSR_i^j = \sum_{t=1}^{T} (y_{it} - x_{it}\ddot{B}_j)^2$, $i = 1, \ldots, N$, and $j = 1, 2$; (b) If $SSR_i^1 \leq SSR_i^2$, individual i is re-assigned to group 1; otherwise, i stays with group 2. Then, (c) update $\{\ddot{I}_1, \ddot{I}_2\}$ and recalculate the fixed effects estimator $(\ddot{B}_1, \ddot{B}_2)$ and SSR_i^j.

3. Re-shuffle individuals unit by unit to form new grouping $\{\ddot{I}_1', \ddot{I}_2'\}$ and calculate $(\ddot{B}_1', \ddot{B}_2')$ and $SSR_i^{j\prime} = \sum_{t=1}^{T} (y_{it} - x_{it}\ddot{B}_j')^2$. If $\sum_j \sum_{i \in I_j'} SSR_i^{j\prime} < \sum_j \sum_{i \in I_j} SSR_i^j$, then repeat 2.(a)–(c) with $(\ddot{B}_1', \ddot{B}_2')$.

[6] There are many variations to the basic algorithm. Harmonic and fuzzy means have also been used instead of simple means. See, for example, Hartigan (1975), Abraham, Cornillion, Matzner-Lober, and Molinari (2003).

Steps 1 to 3 are repeated several times to reduce the effects of the initial group assignment. As discussed in Garcia-Escudero and Gordaliza (1999), the algorithm is known to be sensitive to the presence of outliers. The algorithm can be extended to the situations with $G > 2$ groups.

For i.i.d. data, Pollard (1981) used empirical process arguments to obtain a strong consistency result while Pollard (1982) showed that the centroids estimated by the algorithm are asymptotically normal. However, Pollard (1981) noted that his consistency result does not necessarily apply to algorithms used to find optimal partitions in practice. For example, the algorithm needs to be restarted many times to ensure that the objective function achieves a global minimum. As far as we are aware of, the asymptotic properties of K-means algorithm used in practice (with multiple restarts) is not available. Our panel K-means algorithm suffers from the same caveat.

While our pseudo threshold procedure minimizes the same objective function as K-means, three implementation issues are noteworthy. First, we only estimate the ordered regression once. The K-means algorithm makes random initial guesses of the centroids and then evaluates if a move to a different group is desirable unit by unit. This makes the K-means method computationally costly when N is large. Furthermore, when there are multiple alternatives and N is large, convergence of the K-means can be slow. Second, because we follow the structural break literature and search for the optimal threshold value in the [.1, .9] fraction of the sample, our approach is less sensitive to outliers. Simulations bear this out. Third, we search for the second threshold value after the first threshold value determines two subgroups. In contrast, the K-means is a global procedure. As such, units found to be in Group 1 by the K-means when $G = 2$ can be in Group 2 when $G = 3$.

The K-means method also has two advantages. First, the algorithm relies only on the pooled estimator $\hat{B}_g$ which is $\sqrt{NT}$ consistent, and does not require the individual estimates $\hat{\beta}_i$'s, which are $\sqrt{T}$ consistent. Thus the K-means method will be more precise even when N or T is small. In contrast, the pseudo threshold approach requires N and T to be large. Second, the K-means method considers every unit in the sample for a move to a different group. Our pseudo threshold method moves all those units with $\hat{q}_i$ above and below the threshold value simultaneously. The simultaneous move method is fast, but can be inaccurate when the ordering of $\hat{q}_i$ does not agree with q_i, as may be the case when the sample size is small, or when q_i does not provide complete information about the group structure. We can therefore expect a trade-off between precision and speed in the different methods.

5 Determining G

The objective of our analysis is to estimate the parameters of the panel data model. While group classification is not our main focus, the foregoing analysis assumes that the number of groups G is known. An informal way of determining G is to graph the value of the objective function (in our case, the S_{NT} for a given G) against the number of groups G and then locate the 'knee point' at which the objective function starts to flatten. More formal procedures have been proposed to determine the number of clusters outside of a regression framework. Milligan and Cooper (1985) considered 30 procedures and found that the global procedure of Calinski and J. Harabasz (1974) works best, while the local procedure of Duda and Hart (1973) is second. But as Sugar and James (2003) pointed out, most methods were developed for a specific problem and are somewhat ad-hoc. The statistics literature is still in search of a procedure that can determine the number of groups in a general setting.

Determining the number of clusters shares similarity with determining the number of break points or thresholds. In those problems, we can use a sup-Wald type test for the null hypothesis of no threshold effect.[7] However, there are three features that make the SupW test for parameter homogeneity infeasible here. First, $\hat{B}_1$ and $\hat{B}_2$ are estimated from two split samples ordered by $\hat{\beta}_i$. One sample will have smaller values of $\hat{\beta}_i$ and the other will have the larger values. Thus, the pooled estimate will be biased if $B_1 = B_2$. Second, $\hat{B}_1$ and $\hat{B}_2$ are correlated when $B_1 = B_2$, making inference non-standard. Third, as $\hat{q}_i$ is ordered, bootstrap procedures valid for cross-sectionally independent data are now invalid. Furthermore, determining the number of clusters is not really our ultimate objective of interest.

We experimented with a number of methods developed in the literature, but they tend to be inaccurate unless when the parameters in different groups are very far apart. However, two methods seem promising, both not previously considered in the literature. The first is motivated by a result of Bai (1997) who shows that in time series regressions, the break fractions can be consistently estimated one at a time. We thus consider a sequential test of parameter homogeneity, which can be stated as $H_0 : \beta_i = B \ \forall i$. If we cannot reject H_0, then $G = 1$. If we reject, then we partition the sample into two using PSEUDO1, PSEUDO2 or the K-means. We then test if H_0 holds for each of the subgroups. We may conclude that G is 2 if we cannot reject subsample heterogeneity.

[7] See, for example, Davies (1977), Andrews and Ploberger (1994), Hansen (1996), Bai (1997), and Caner and Hansen (2004).

If subsample homogeneity is rejected, the sample is split again until the null hypothesis cannot be rejected for the subsamples.

To implement this sequential procedure, we use the dispersion t test for the null hypothesis of parameter homogeneity proposed by Pesaran and Yamagata (2008). It is defined as

$$t_g = \frac{\sqrt{N}(\xi_N/N - K)}{\sqrt{2KG}}, \tag{9}$$

where K denotes the number of the regressors, $\xi_N = \sum_{i=1}^{N} \widetilde{\sigma}_i^{-2}(\hat{\beta}_i - \widetilde{B}_w)' \left(\sum_{t=1}^{T} x_{it}' x_{it}\right) (\hat{\beta}_i - \widetilde{B}_w)'$, $\widetilde{B}_w$ is the weighted pooled fixed effects estimator of Swamy (1970), and $\widetilde{\sigma}_i^2$ is obtained by fixed effects estimation of B under the null hypothesis of homogeneity. This test allows for heteroskedasticity and non-normally distributed errors and is consistent as N and T go to infinity jointly such that $\sqrt{N}/T^2 \to 0$.

The second method for determining G proceeds along the lines of Bai and Perron (1998) for determining multiple structural breaks. As we have panel data, the BIC for g groups is

$$BIC(g) = \log\left(\Sigma_{NT}(g, \hat{\gamma}, \hat{q})\right) + gK \cdot \frac{c_{NT}\log(NT)}{NT} + (g-1)\frac{\log(N^2)}{N^2}, \tag{10}$$

where

$$\Sigma_{NT}(g, \hat{\gamma}, \hat{q}) = \frac{1}{G}\sum_{k=1}^{G} \frac{1}{N_g T} \sum_{i \in \hat{I}_g} \sum_{t=1}^{T} (y_{it} - x_{it}\hat{B}_g(\hat{\gamma}))^2.$$

The goodness of fit component of the BIC is computed as the average (over groups) of the regression error variance in each group, where group membership is determined by either PSEUDO1, PSEUDO2, or K-means. The penalty of $\log(NT)/NT$ is guided by the fact that $\hat{B}_w$ is $\sqrt{NT}$ consistent under the null. In this case, the BIC should consistently select g if $\frac{c_{NT}^*}{NT} \to 0$ but $c_{NT}^* \to \infty$ as $N, T \to \infty$. When all regressors and γ are observed, BIC obtains with $c_{NT}^* = \log(NT)$, or $c_{NT} = 1$. We consider a heavier penalty as $\hat{q}_i$ and $\hat{\gamma}$ are themselves estimated. Base on extensive simulations, we set the penalty on additional regressors as $c_{NT} = \sqrt{\min[N, T]}$. As $c_{NT}^* = c_{NT}\log(NT)$ diverges, and $\frac{c_{NT}^*}{NT} \to 0$, the required conditions for consistent model selection are satisfied. Furthermore, the breakpoint literature suggests $\hat{\gamma}$ is super-consistent with variance that vanishes at rate N^2. We put a penalty of $\log(N^2)$ on each threshold variable, giving an overall penalty on $\hat{\gamma}$ of $(g-1)\log(N^2)/N^2$.

As discussed earlier, our clustering method depends on the choice of G, but a $\hat{G}$ that equals the correct G need not produce pooled estimates that are closest to the true group parameters. More parameters are estimated as more groups are allowed. The increased sampling variability must

be balanced against the bias induced by pooling units from different groups. Indeed, Baltagi and Griffin (1997) and Baltagi, Griffin, and Xiong (2000) find that models with complete heterogeneity yield inferior predictions than those that impose homogeneity even though they find substantial heterogeneity in the price elasticity of demand for gasoline. Thus, while $\hat{G}$ is an interesting result in its own right, accuracy in selecting G may not reflect how well a model captures parameter heterogeneity. Thus in the simulations to follow, we evaluate both the accuracy of the estimated parameters and of G.

6 Simulations and Applications

We now use Monte Carlo simulations to examine the finite sample properties of the methods considered. We generate data as follows. For $G = 2, 3$, $K = 1, 2$, $t = 1, \ldots, T$, and $i = 1, \ldots, N$,

$$\widetilde{y}_{it} \quad = \quad \alpha_i + \sum_{g=1}^{G} \left(\sum_{k=1}^{K} \widetilde{x}_{k,it} B_{gk} \right) 1(i \in I_g^0) + \widetilde{e}_{it},$$

where $\alpha_i \sim$ i.i.d. $N(1,1)$, $\widetilde{x}_{k,it} \sim$ i.i.d. $N(1,3)$, and independent of $\widetilde{e}_{it}$, $\widetilde{e}_{it} \sim N(0,1)$ is i.i.d. over i and t. When $G = 2$, we randomly assign individuals into two groups $\{I_1^0, I_2^0\}$ with size $N_1^0 = \lfloor 2N/3 \rfloor$ and $N_2^0 = N - N_1^0$, where $\lfloor A \rfloor$ denotes the maximum integer that does not exceed real number A. When $G = 3$, we randomly assign individuals into three groups $\{I_1^0, I_2^0, I_3^0\}$ with size $N_k^0 = \lfloor N/3 \rfloor$ for $k = 1, 2$, and $N_3^2 = N - N_1^0 - N_2^0$.

We consider the following configurations:

i For $(G, K) = (2, 1)$, $B_1 = 0.3$ and $B_2 = 0.9$.

ii For $(G, K) = (3, 1)$, $(B_1, B_2, B_3) = (0.3, 0.5, 0.8)$.

iii For $(G, K) = (2, 2)$, $B_1 = (B_{11}, B_{12})' = (0.1, 0.3)'$ and $B_2 = (B_{21}, B_{22})' = (2/3, 0.6)'$.

iv For $(G, K) = (3, 2)$, $B_1 = (B_{11}, B_{12})' = (0.3, -0.3)'$, $B_2 = (B_{21}, B_{22})' = (0.5, 0)'$ and $B_3 = (B_{31}, B_{32})' = (0.7, 0.3)'$.

These parameterizations give an R^2 of around 0.5. We consider combinations of (N, T) with $N = (50, 100, 200, 500)$ and $T = (20, 50, 100, 200)$. Throughout, we use $M = 1000$ replications, holding $\{I_1^0, \cdots, I_K^0\}$ and α_i fixed over replications.

We consider three clustering methods: PSEUDO$_1$ with $\hat{q}_i = \hat{\beta}_i - \hat{B}_w$ or PSEUDO$_2$ with $\hat{q}_i = \hat{\beta}_i$, and the K-means. To estimate G by t_g defined in (9) or $BIC(g)$ defined in (10), we need pooled

estimation for each group which in turn depends on which units are in the group. As such, the t_g can be implemented in conjunction with any of the three clustering methods. Because the t-test is asymptotically standard normal, we use the critical value of 1.96. To reduce the effect of the initial group assignment for the K-means method, for each group size $\dot{G} = 1, \ldots, G_{\max} = 4$ under consideration, we randomly draw $\dot{G} + 5$ sets of initial assignment and take the results from the set with the minimum sum of squared residuals. When $K > 1$, the threshold variable is determined according to (8). In order to avoid having groups with too few units (which arises primarily when N or T is small), we restrict the number units in each group to be at least $\max\{10, 0.1N\}$.

Depending on how far the coefficients in two groups are separated, a model estimated with homogeneity imposed might well approximate the conditional mean better, even though misclssification rate would be high. Comparing goodness of fit when the number of free parameters differs across models will always favor a complex model. Instead, we record the root mean squared error of the estimates (RMSE) defined as

$$RMSE = \frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{N} \frac{1}{N} \left\| \hat{B}_i^{(m)}(\hat{G}) - \beta_i(G) \right\|^2$$

where $B_i^{(m)}(\hat{G})$ is the pooled slope parameter in the m-th replication estimated for the ith unit based on $\hat{G}$ and group assignment determined by one of the three clustering methods, and $\beta_i(G)$ is the true slope coefficient for the unit. Figure 1 plots the RMSE for the different configurations of the sample size. For each configuration, we have results (starting from the leftmost bar) for K-means, PSEUDO1, PSEUDO2 using the t_g test to determine G. This is followed by K-means, PSEUDO1, PSEUDO2 using the BIC to determine G. To study the error due to miss-classification and/or incorrect estimation of G, we also graph (i) the RMSE when G and group membership are known, and (ii) when G is known but group membership is not and is determined by PSEUDO1 or the K-means.

As can be seen from Figure 1, regardless of the method used to cluster the sample (which can be K-means, PSEUDO1, or PSEUDO2), the RMSE tends to decrease as N or T increases, but an increase in T has a larger impact on RMSE than an increase in N. Conditional on $\hat{G}$, K-means and PSEUDO1 have similar RMSEs when T is large, but the K-means has smaller errors when T is small. This is to be expected since the pseudo threshold method requires $\sqrt{T}$ consistent estimation of the individual slope parameters. However, PSEUDO2 is much more sensitive to N and T and is clearly inferior to PSEUDO1 and K-means. Further investigation reviews that the issue is not with

using $\hat{\beta}_i$ as threshold variable, but with using $\hat{B}_\omega$ as threshold value. If we let the data determine γ, PSEUDO2 is similar to PSEUDO1. In other words, even though $\hat{\gamma} = \hat{B}_\omega$ is a valid threshold value, it is inferior to estimating γ by minimizing $S_{NT}(\gamma, \hat{q})$, As this is also the case for other DGPs, results for PSEUDO2 will not be discussed further.

While $\hat{B}_i(\hat{G}) - \beta_i(G)$ tells us how well we have captured the heterogeneity in coefficients, this metric has no obvious relation to how far $\hat{G}$ is from G. Define the accuracy of selecting the true group number GR by

$$GR = \frac{1}{M}\sum_{m=1}^{M} 1(\hat{G}^m = G).$$

Also define the accuracy of classification by

$$CR = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{\hat{G}} 1\left((i \in \hat{I}_g^{(m)})\bigcap(i \in I_g^0)\right).$$

Note that GR evaluates difference between $\hat{G}$ and G while CR indicates accuracy of assigning individuals into groups. Thus these two indicators need not be close.

The results for GR and CR are presented in Figures 2 and 3 respectively. Observe first that while the BIC accurately selects G when $T \leq N$, the GR is quite low when $N > T$. The t_g test is more robust to variations in the sample size, regardless of the clustering method used. Not surprisingly, when N and T are small, and there are more than two groups, G cannot be estimated precisely. The K-means used in conjunction with t_g selects the correct G wit probability around 0.95 when $T \geq 50$, which is higher than the t_g used in conjunction with PSEUDO1. The results when $K = 1$ are similar to those when $K > 1$. Because we can obtain additional information from the second regressor, the estimated G with two regressors tend to be more accurate than in the one regressor case. The CR evidently improves as T increases and does not change much with N as Theorem 2 suggests.

The RMSE presented in Figure 1 is based on the slope parameters. To give an overall sense of goodness of fit, we consider out-of-sample validation as follows. For units $i' = 1, \ldots, J = \lfloor N/3 \rfloor$ not in the estimation sample, where $\lfloor A \rfloor$ the largest integral not exceeding A, we obtain individual estimates of their slope parameters, denoted $\hat{\beta}_{i'}$. We then use the threshold of $\hat{\gamma}$ to assign the unit into one of the $\hat{G}$ groups. Pooled estimates of the slope parameters are then obtained and the sum of squared residuals $S_{JT}(\hat{\gamma}, \hat{q}, \hat{G})$ for the J units is recorded. For comparison, we consider three benchmarks: (i) $S_{JT}(G = 1)$ for the model that imposes homogeneity, (ii) $S_{JT}(\hat{\gamma}, \hat{q}, G)$ for the case

when G is known but group membership is not, and (iii) $S_{JT}(\gamma, q, G)$ for the case when both G and group membership are known. As we can see from Figure 4 the sum of squared residuals with unknown group membership and unknown G are similar to those when G and/or group membership has to be estimated. This suggests that the clustering error that our methods produce has little impact of the fitted regression model. Importantly, the sum of squared residuals is always smaller than wrongly imposing homogeneity.

We also generate data from a dynamic panel model with group specific parameters:

$$\widetilde{y}_{it} = \alpha_i + \rho_g \widetilde{y}_{i,t-1} + \phi_g t + \widetilde{e}_{it}, \quad \text{if } i \in I_g^0. \tag{11}$$

We set $\alpha_i = 0$ for all i's, $(\rho_1, \rho_2) = (.3, .8)$, $(\phi_1, \phi_2) = (0, .03)$, $\widetilde{e}_{it} \sim N(0, 1)$ is i.i.d. over i and t. The results are presented in Table 1. As in the static DGP, the RMSE tends to decrease as T increases, and the RMSE is smaller when G is chosen correctly. When T is small, the t_g is preferred over the BIC judged in terms of both the CR and GR. The RMSE for PSEUDO1 and K-means are similar when T is large.

Overall, the results lead to four conclusions. First, the t_g-test in conjunction with K-means or PSEUDO can accurately estimate the number of groups in our setup. Second, PSEUDO2 is inferior to PSEUDO1. Third, existing methods in the literature fail to select G accurately. The proposed method of sequential testing using t_g is best, while the BIC with an additional penalty term is reasonably accurate when T is large. Fourth, for small sample size, the K-means is much preferred over PSEUDO1. However, when the sample size is large, PSEUDO1 is as effective as K-means. In this latter case, PSEUDO1 has a distinct computational advantage as the number of regressions under the threshold approach is of order N, while the K-means involves an enumeration of G^N regressions.[8]

6.1 Empirical Study

The existence of "convergence clubs" has generated much research interests in the growth literature. A group of countries with a similar steady state that can be characterized by the same linear model are said to form a convergence club. Most studies use observed variables to group the countries and then estimate the group specific parameters. See Durlauf, Kourtellos, and Tan (2008) for a survey. Some find that the quality of institutions and ethnic fractionalization are the most important

[8] In unreported results, ignoring the individual-specific fixed effects leads to more inaccurate estimates of the clusters, as expected.

determinants of economic growth. Others argue that the savings rate is more important, as are education-related variables. See Barro and Sala-i-Martin (2003) for a discussion on issues relating to empirical growth regressions.

To motivate the estimation issue when group membership is not known, consider the model used in Lee, Pesaran, and Smith (1997) for 69 countries, taken from the PWT v6.2 by Heston, Summers, and Aten (2006) for the sample 1965 to 2003.[9] The regression model is

$$\widetilde{y}_{it} = \alpha_i + \rho_g \widetilde{y}_{it-1} + \phi_g t + \widetilde{e}_{it}, \quad i \in I_g, \quad g = 1, 2, \ldots, G, \tag{12}$$

where $\widetilde{y}_{it}$ is the log per-capita output, α_i denotes country-specific fixed effects, ρ_g and ϕ_g, $g = 1, \ldots, G$, $G = \{1, 2, \ldots, G_{\max}\}$, are group specific. While previous studies allow for differences in α_i and in estimation of ϕ_i, heterogeneity in ρ rarely allows.

We first obtain, for each i, estimates of ϕ_i and ρ_i from individual time series regression. The t_g test suggests either $G = 4$ or $G = 5$ depending on whether K-means or PSEUDO1 is used. We choose the more parsimonious specification of $G = 4$. The results, reported in the top panel of Table 2, show that the groups have different features in terms of $\hat{\rho}_g$ and $\hat{\phi}_g$. Specifically, $\hat{\rho}_g$ is much smaller in groups 1 and 3 than in groups 2 and 4. Furthermore, $\hat{\phi}_g$ is negative in groups 1 and 2, but positive in groups 3 and 4. Both ρ_g and ϕ_g are thus heterogeneous across groups. Interestingly, while the 21 OECD countries do not all belong to the same group, the fast growing countries like Indonesia, Korea, Malaysia, and Thailand are in the same (non-OECD) group. A priori information would unlikely arrive at such a grouping.

Equation (12) assumes cross-section independence in $\widetilde{e}_{it}$. Pesaran (2006) suggests controlling for cross-correlated errors by adding the cross-section average of appropriate variables to the pooled regression. Our analysis consists of both pooled and individual regressions. To guard against simultaneity bias, we add $\Delta\bar{y}_{t-1}$ and $\Delta\bar{y}_{t-2}$ to both the pooled and the individual regressors, where $\bar{y}_s = N^{-1}\sum_i^N \widetilde{y}_{is}$, $\bar{\triangle} y_s = \bar{y}_s - \bar{y}_{s-1}$, $s = 1, \ldots, T$. The results, reported in the bottom panel of Table 2, show that after allowing for cross-section dependence, ϕ_g in group 1 is now positive, and group 3 may warrant further splitting. However, the parameter estimates reinforce the main finding that income dynamics across countries differ in two dimension: in the growth rate and in the speed of adjustment to equilibrium.

[9] See Mankiw, Romer, and Weil (1992) on how to select 75 intermediate countries. However, Germany is removed from data set due to consolidation. Due to limitation of data, we also remove Bangladesh, Bolivia, Botswana, Haiti, and Myanmar.

7 Conclusion

We use time series estimates of the coefficients for each unit to form 'pseudo threshold variables'. These are then used to partition the panel into groups. Our model based method is shown to consistently estimate the true number of groups identified by distinct coefficients on the covariates. The methodology can be modified to use weighted least squares with weight $1/\hat{\sigma}_i$. Simulations show that taking into account of heteroskedasticity when this feature is present in the data yields smaller RMSEs in the parameter estimates.

APPENDIX: PROOFS

To prove Theorem 1, we let $I^0 = (I_1^0, I_2^0)$ be the true group membership and let $I = (I_1, I_2)$ denote group membership other than (I_1^0, I_2^0). Suppose that the DGP is

$$\begin{aligned} \widetilde{y}_{it} &= \alpha_i + \widetilde{x}_{it}B_1 + \widetilde{e}_{it}, \text{ for } i \in I_1^0, \\ \widetilde{y}_{it} &= \alpha_i + \widetilde{x}_{it}B_2 + \widetilde{e}_{it}, \text{ for } i \in I_2^0. \end{aligned}$$

We will consider the general case where $B_2 - B_1 = \nu T^{-\alpha}$, $0 \leq \alpha < 1/2$, ν does not depend on T, and $||\nu|| > 0$. Then $\alpha = 0$ corresponds to the case when $B_2 - B_1 = \nu \neq 0$.

For $j, k = 1, 2$, let N_{kj} be the number of individuals assigned to be in group j by $I = (I_1, I_2)$ when individuals truly belong to group k and let $\hat{B}_{kj}$ denote the estimator of slope parameter for $i \in I_k^0 \cap I_j$, $\hat{B}_j$ for $i \in I_j$, and $\hat{B}_k^0$ for $i \in I_k^0$. Let N_k^0 denote the number of individuals truly belonging to group k and let $N_1 = N_{11} + N_{21}$ and $N_2 = N_{22} + N_{12}$. Notice that in Theorem 1 individuals are ordered by q_i. Without loss of generality, we assume $N_{12} = 0$. Therefore, $N_1 = N_{11} + N_{21}$, $N_2 = N_{22}$ and we let $N_s = N_s(I, I^0) = N_{21}$ be the number of misclassified units. We then define $\mathcal{I}(C) = \{I : N_s(I, I^0) \geq C/T^{1-2\alpha}\}$.

Let $z_{it} = \widetilde{z}_{it} - \overline{\widetilde{z}}_i$, where $\widetilde{z}_{it}$ can be $\widetilde{y}_{it}, \widetilde{x}_{it}, \widetilde{e}_{it}, \widehat{\widetilde{e}}_{it}$, and $\overline{\widetilde{z}}_i = \frac{1}{T}\sum_{t=1}^T \widetilde{z}_{it}$. Then for $(k, j) = (1, 1), (2, 1)$, and $(2, 2)$,

$$y_{it} = x_{it}B_k + e_{it} = x_{it}\hat{B}_j + \hat{e}_{it}^{kj},$$

where $\hat{e}_{it}^{kj} = e_{it} + x_{it}(B_k - \hat{B}_j)$. Thus,

$$\hat{e}_{it}^{kj2} = e_{it}^2 + (\hat{B}_j - B_k)'x_{it}'x_{it}(\hat{B}_j - B_k) + 2e_{it}[x_{it}(B_k - \hat{B}_j)].$$

It is convenient to define $x_i = (x_{i1}', \ldots, x_{iT}')'$, $e_i = (e_{i1}, \ldots, e_{iT})'$, and

$$H_{kj} = \sum_{i \in I_k^0 \cap I_j} x_i'x_i \text{ with } H_1 = H_{21} + H_{11} \text{ and } H_2 = H_{22}.$$

Also, we define $\Delta_{kj} = \hat{B}_{kj} - B_k = H_{kj}^{-1}\sum_{i \in I_k^0 \cap I_j} x_i'e_i$ and $\Delta_k^0 = \hat{B}_k^0 - B_k$ for $(k, j) = (1, 1), (2, 1)$ or $(2, 2)$. Let $S_{NT}(I)$ and $S_{NT}(I^0)$ denote the total sum of squared residuals under I and I^0, respectively. After some tedious algebra, we obtain

$$S_{NT}(I) - S_{NT}(I^0) = \psi_1 + \psi_2 + \psi_3, \tag{13}$$

where

$$\begin{aligned} \psi_1 &= T^{-\alpha}\nu'(H_{21}H_1^{-1}H_{11})\nu T^{-\alpha} \\ \psi_2 &= -2T^{-\alpha}\nu'(H_{21}H_1^{-1}H_{11}\Delta_{11}) + 2T^{-\alpha}\nu'(H_{11}H_1^{-1}H_{21}\Delta_{21}) \\ \psi_3 &= -(H_{11}\Delta_{11} + H_{21}\Delta_{21})'H_1^{-1}(H_{11}\Delta_{11} + H_{21}\Delta_{21}) + \Delta_1^{0\prime}H_{11}\Delta_1^0 - \Delta_{22}'H_{22}\Delta_{22} + \Delta_2^{0\prime}(H_{22} + H_{21})\Delta_2^0. \end{aligned}$$

The detailed proof of equation (13) is available from authors upon request. In the following, we will show a few lemmas used in Theorem 1.

Lemma A1 *Under Assumptions 1 and 3,*

(a) For each i, $T^{-1/2}x_i'e_i \xrightarrow{d} N(0, \sigma_i^2 Q_i)$ as $T \to \infty$.

(b) Let I_* denote a nonempty subset of the whole sample and let $N_* \geq 1$ be the number of units in I_*. Then, $(N_*T)^{-1/2}\sum_{i \in I_*} x_i'e_i \xrightarrow{d} N(0, Q_*)$ as $(N_*, T) \to \infty$ jointly, where $Q_* = \lim_{N_* \to \infty} N_*^{-1}\sum_{i \in I_*} \sigma_i^2 Q_i$.

Proof of Lemma A1: Lemma A1(a) follows from Assumption 1 by central limit theorems used in classical linear regression models. Lemma A1(b) directly follows because $\widetilde{e}_{it}$ is assumed to be cross-sectionally independent and $\mathrm{E}||Q_{iT}||$ is finite by Assumption 3(a). See Lemma 4 in Pesaran (2006).

Lemma A1' *Under Assumptions D1–D3 with* $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$,

(a) For each i, $T^{-1}x_i'x_i = O_p(1)$, $0 < \mathrm{E}|T^{-1}x_i'x_i|$ *is finite, and* $T^{-1}x_i'x_i$ *is strictly positive. Also,* $T^{-1/2}x_i'e_i \xrightarrow{d} N(0, \sigma_i^2 Q_i)$ *as* $T \to \infty$.

(b) Let I_* *denote a nonempty subset of the whole sample and let* $N_* \geq 1$ *be the number of units in* I_*. *Then,* $(N_*T)^{-1}\sum_{i\in I_*} x_i'x_i = O_p(1)$ *and* $(N_*T)^{-1/2}\sum_{i\in I_*} x_i'e_i = O_p(1)$.

Proof of Lemma A1': The first part of Lemma A1(a) follows from the fact that $0 < \mathrm{E}|x_i'x_i/T| = \sigma_i^2/(1-\beta_i^2) < \infty$, $x_i'x_i/T = \sigma_i^2/(1-\beta_i^2) + O_p(T^{-1/2})$, and $x_i'x_i/T > 0$. The second part of Lemma A1'(a) follows from Assumptions D1 and D3 by central limit theorems used in a stationary AR(1) regression model. Next, consider Lemma A1'(b). Following the proof of Theorem 3 in Pesaran and Yamagata (2008, p.84), we obtain

$$(N_*T)^{-1}\sum_{i\in I_*} x_i'x_i = N_*^{-1}\left(\sum_{i\in I_*\cap I_1^0}\frac{\sigma_i^2}{1-B_1^2} + \sum_{i\in I_*\cap I_2^0}\frac{\sigma_i^2}{1-B_2^2}\right) + O(T^{-1}) + O_p(N_*^{-1/2}T^{-1/2}), \quad (14)$$

$$(N_*T)^{-1/2}\sum_{i\in I_*} x_i'e_i - \mathrm{Bias}_{N_*T}(B_1,B_2) \xrightarrow{d} N(0, Q_*),$$

where

$$\mathrm{Bias}_{N_*T} = -\lim_{N_*,T\to\infty \text{ jointyly}} \sqrt{\frac{N_*}{T}} N_*^{-1}\left[\left(\sum_{i\in I_*\cap I_1^0}\frac{\sigma_i^2}{1-B_1} + \sum_{i\in I_*\cap I_2^0}\frac{\sigma_i^2}{1-B_2}\right)\right], \quad (15)$$

$$Q_* = \lim_{N_*\to\infty} N_*^{-1}\left(\sum_{i\in I_*\cap I_1^0}\frac{\sigma_i^4}{(1-B_1^2)} + \sum_{i\in I_*\cap I_2^0}\frac{\sigma_i^4}{(1-B_2^2)}\right).$$

Since $N_* \leq N$ and $0 \leq N/T < \infty$ as $N, T \to \infty$ jointly, $(N_*T)^{-1}\sum_{i\in I_*} x_i'x_i = O_p(1)$ and $(N_*T)^{-1/2}\sum_{i\in I_*} x_i'e_i = O_p(1)$. □.

Lemma A2 *Under Assumptions 1–3 or under Assumptions D1–D3 with* $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$,

(a) For $(k,j) = (1,1),(2,1)$ *and* $(2,2)$, *if* $N_{kj} \geq 1$, $H_{kj}/(N_{kj}T) = O_p(1)$ *and* $H_{kj}\Delta_{kj}/\sqrt{N_{kj}T} = O_p(1)$. *Otherwise,* $H_{kj} = 0$ *and* $H_{kj}\Delta_{kj} = 0$.

(b) For $k = 1,2$, $H_k^0/(N_k^0T) = O_p(1)$ *and* $H_k^0\Delta_k^0/\sqrt{N_k^0T} = O_p(1)$, *where* $H_1^0 = H_{11}$ *and* $H_2^0 = H_{22} + H_{21}$.

(c) For $j = 1,2$, $H_j/(N_jT) = O_p(1)$, $(H_{11}\Delta_{11} + H_{21}\Delta_{21})/\sqrt{N_1T} = O_p(1)$, *and* $H_{22}\Delta_{22}/\sqrt{N_2T} = O_p(1)$.

Proof of Lemma A2: We first show the proof under Assumptions 1–3. Consider (a). When $N_{kj} > 0$, $H_{kj}/(N_{kj}T) = O_p(1)$ directly follows from Assumption 3. Since $\Delta_{kj} = H_{kj}^{-1} \sum_{i \in I_k^0 \cap I_j} x_i' e_i$ and by Lemma A1,

$$N_{kj}^{-1/2} T^{-1/2} H_{kj} \Delta_{kj} = H_{kj} H_{kj}^{-1} \left(N_{kj}^{-1/2} T^{-1/2} \right) \sum_{i \in I_k^0 \cap I_j} x_i' e_i = O_p(1).$$

When $I_k^0 \cap I_j = \emptyset$, $N_{kj} = 0$. It follows that $H_{kj} = 0$ and $\sum_{i \in I_k^0 \cap I_j} x_i' e_i = 0$. Analogously, (b) holds by Assumptions 1–3, Lemma A1, and the fact that $H_k^0 \Delta_k^0 = \sum_{i \in I_k^0} x_i' e_i$. For (c), the first claim holds because $H_1 = H_{11} + H_{21}$, $H_2 = H_{22}$, $N_1 = N_{11} + N_{21}$, and $N_2 = N_{22}$. Because $(H_{11}\Delta_{11} + H_{21}\Delta_{21}) = \sum_{i \in I_1} x_i' e_i$ and $H_{22}\Delta_{22} = \sum_{i \in I_2} x_i' e_i$, the second claim follows by Lemma A1.

Similarly, under the Assumptions D1–D3, the same results follows by Lemma A1' in a dynamic model with $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$.

□.

Lemma A3 *Under Assumptions 1–3 or under Assumptions D1–D3 with* $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$, $\inf_{I \in \mathcal{I}(C)} N_s^{-1}(T^{-1+2\alpha})\psi_1 > 0$.

Proof of Lemma A3: Notice that

$$(N_s^{-1}T^{-1+2\alpha})\psi_1 = \nu' T^{-1} N_s^{-1} (H_{21} H_1^{-1} H_{11}) \nu. \tag{16}$$

Lemma A3 follows if we can show that the minimal eigenvalue of $T^{-1}N_s^{-1}(H_{21}H_1^{-1}H_{11})$ is bounded away from zero uniformly in $\mathcal{I}(C)$. Notice that $N_s = N_{21}$, $N_1 = N_{11} + N_{21}$, and

$$T^{-1}N_s^{-1}H_{21}H_1^{-1}H_{11} = \frac{H_{21}}{N_{21}T} \left(\frac{H_1}{N_1 T} \right)^{-1} \frac{H_{11}}{N_{11}T} \frac{N_{11}}{N_1}.$$

First, we consider a panel data model under Assumptions 1–3. By Assumption 3, $\frac{H_{21}}{N_{21}T}$ has the minimal eigenvalue bounded away from zero, which is then equivalent to show (16) is strictly positive.

Similarly, consider a dynamic panel model with $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$. Under Assumptions D1–D3, $\frac{H_{21}}{N_{21}T}$, $\frac{H_1}{N_1 T}$, and $\frac{H_{11}}{N_{11}T}\frac{N_{11}}{N_1}$ are all strictly positive. Therefore, the desired results follows. □

Lemma A4 *Under Assumptions 1–3 or under Assumptions D1–D3 with* $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$,
(a) $(N_s^{-1}T^{-1+2\alpha})\psi_2 = o_p(1)$ *uniformly in* $\mathcal{I}(C)$.
(b) $(N_s^{-1}T^{-1+2\alpha})\psi_3 = o_p(1)$ *uniformly in* $\mathcal{I}(C)$.

Proof of Lemma A4: Consider (a) first. Notice that

$$(N_s^{-1}T^{-1+2\alpha})\psi_2 = -2\nu' T^{-1/2+\alpha} \left[N_s^{-1} T^{-1/2} \left(H_{21} H_1^{-1} H_{11} \Delta_{11} - H_{11} H_1^{-1} H_{21} \Delta_{21} \right) \right].$$

When $N_{21} = 0$, $H_{21}H_1^{-1}H_{11}\Delta_{11} - H_{11}H_1^{-1}H_{21}\Delta_{21} = 0$. When $N_{21} > 0$, $\left(N_s^{-1}T^{-1/2}H_{21}H_1^{-1}H_{11}\Delta_{11}\right) = N^{-1/2}O_p(1)$ and $\left(N_s^{-1}T^{-1/2}H_{11}H_1^{-1}H_{21}\Delta_{21}\right) = N_{21}^{-1/2}O_p(1)$ by Lemma A2. Because $-1/2 + \alpha < 0$,

$$-2\nu' T^{-1/2+\alpha} \left[N_s^{-1} T^{-1/2} \left(H_{21} H_1^{-1} H_{11} \Delta_{11} - H_{11} H_1^{-1} H_{21} \Delta_{21} \right) \right] = o_p(1)$$

uniformly in $\mathcal{I}(C)$.

Next consider (b). By Lemma A2(c), $(H_{11}\Delta_{11}+H_{21}\Delta_{21})'H_1^{-1}(H_{11}\Delta_{11}+H_{21}\Delta_{21})=O_p(1)$ and $\Delta_{22}'H_{22}\Delta_{22}=O_p(1)$ uniformly in $\mathcal{I}(C)$. Also, by Lemma A2(b), $\Delta_1^{0\prime}H_{11}\Delta_1^0+\Delta_2^{0\prime}(H_{22}+H_{21})\Delta_2^0=O_p(1)$ uniformly in $\mathcal{I}(C)$. Thus,

$$(N_s^{-1}T^{-1+2\alpha})\psi_3=(N_s^{-1}T^{-1+2\alpha})(H_{11}\Delta_{11}+H_{21}\Delta_{21})'H_1^{-1}(H_{11}\Delta_{11}+H_{21}\Delta_{21})$$
$$+(N_s^{-1}T^{-1+2\alpha})\Delta_{22}'H_{22}\Delta_{22}+(N_s^{-1}T^{-1+2\alpha})\left[\Delta_1^{0\prime}H_{11}\Delta_1^0+\Delta_2^{0\prime}(H_{22}+H_{21})\Delta_2^0\right]=o_p(1).$$

Proof of Theorem 1: We first show $N_s=O_p(T^{-1+2\alpha})$. Define $\mathcal{I}(C)=\{I: N_s(I,I^0)\geq C/T^{1-2\alpha}\}$. We want to show that for any ε and $C>0$, $P(N_s>\frac{C}{T^{1-2\alpha}})<\varepsilon$. Since $P(N_s>\frac{C}{T^{1-2\alpha}})<P(I\in\mathcal{I}(C))$, by the definition of I, it suffices to show that

$$P\left(\sup_{I\in\mathcal{I}(C)}SSR(I^0)\geq SSR(I)\right)<\varepsilon.$$

Notice that

$$P\left(\sup_{I\in\mathcal{I}(C)}SSR(I^0)\geq SSR(I)\right)\leq P\left(\sup_{I\in\mathcal{I}(C)}\frac{-(\psi_2+\psi_3)}{N_sT^{1-2\alpha}}\geq\inf_{I\in\mathcal{I}(C)}\frac{\psi_1}{N_sT^{1-2\alpha}}\right).$$

By Lemma A3, $\inf_{I\in\mathcal{I}(C)}\psi_1/(N_sT^{1-2\alpha})$ is strictly positive. By Lemma A4: $\sup_{I\in\mathcal{I}(C)}-(\psi_2+\psi_3)/(N_sT^{1-2\alpha})=o_p(1)$. Together with these lemmas,

$$P\left(\sup_{I\in\mathcal{I}(C)}\frac{-(\psi_2+\psi_3)}{N_sT^{1-2\alpha}}\geq\inf_{I\in\mathcal{I}(C)}\frac{\psi_1}{N_sT^{1-2\alpha}}\right)<\varepsilon,$$

and, therefore, $N_s=O_p(T^{-1+2\alpha})$. Since $N_s/N=O_p(N^{-1}T^{-1+2\alpha})$, $\hat{B}_j\to B_j$ and Theorem 1 follows. □

Proof of Lemma 1: Recall that $\omega=(\sum_{i=1}^N\hat{Q}_i)^{-1}\sum_{i\in I_1^0}\hat{Q}_i$. Let $\hat{B}_j^0=(\sum_{i\in I_j^0}x_i'x_i)^{-1}\sum_{i\in I_j^0}x_i'y_i$. By direct calculations,

$$\begin{aligned}
&\sqrt{NT}\left[\hat{B}_\omega-(\omega B_1+(1-\omega)B_2)\right]\\
=\;&\sqrt{NT}\left[\left(\sum_{i=1}^N x_i'x_i\right)^{-1}\left(\sum_{i\in I_1^0}x_i'y_i+\sum_{i\in I_2^0}x_i'y_i\right)-(\omega B_1+(1-\omega)B_2)\right]\\
=\;&\sqrt{NT}\omega\left(\left(\sum_{i\in I_1^0}x_i'x_i\right)^{-1}\sum_{i\in I_1^0}x_i'y_i-B_1\right)+\sqrt{NT}(1-\omega)\left(\left(\sum_{i\in I_2^0}x_i'x_i\right)^{-1}\sum_{i\in I_2^0}x_i'y_i-B_2\right)\\
=\;&\sqrt{\frac{N}{N_1^0}}\left[\omega\sqrt{N_1^0T}\left(\hat{B}_1^0-B_1\right)\right]+\sqrt{\frac{N}{N_2^0}}\left[(1-\omega)\sqrt{N_2^0T}\left(\hat{B}_2^0-B_2\right)\right]+\kappa_{NT}+o_p(1),
\end{aligned}$$

where $\kappa_{NT}=0$ under Assumptions 1–3, and under Assumptions D1-D3, $\kappa_{NT}=O_p(1)$ by Lemma A1'(b) and $\sum_{i\in I_g^0}x_i'x_i/N_g^0T>0$, $g=1,2$. Thus,

$$\sqrt{NT}\left[\hat{B}_\omega-(\omega B_1+(1-\omega)B_2)\right]=O_p(1),$$

and $\hat{B}_\omega$ is consistent for $\omega B_1+(1-\omega)B_2$. Also, if $\omega_0=\text{plim}_{N\to\infty}\omega$, then $\hat{B}_\omega$ is consistent for $\omega_0B_1+(1-\omega_0)B_2$. □

Proof of Theorem 2: If $N_s/N \to 0$, then we have $\hat{B}_j(\hat{q}) \to B_j$. In the following, we will show $N_s/N \to 0$ under PSEUDO1 and PSEUDO2, respectively.

Consider PSEUDO1 with $B_2 - B_1 = \nu T^{-\alpha}, 0 \leq \alpha < 1/2$ first. Let Γ^0 be a set of threshold values that will achieve correct clustering. Let $\gamma^0_{\min} = \min_\gamma\{\gamma : \gamma \in \Gamma^0\}$ and $\gamma^0_{\max} = \max_\gamma\{\gamma : \gamma \in \Gamma^0\}$. Then for any $\gamma^0 \in [\gamma^0_{\min}, \gamma^0_{\max}]$,

$$F(\gamma^0) = P(q_i^0 < \gamma^0) = \frac{\sum_{i=1}^N 1(q_i^0 < \gamma^0_{\max})}{N} = \frac{N_1^0}{N}.$$

Consider the following cases:

I. Γ^0 known, q_i estimated Suppose we know Γ^0 but not q_i^0. Let $\frac{\hat{c}_i}{\sqrt{T}} = \hat{q}_i - q_i^0$

$$\begin{aligned}
\frac{N_s}{N} &= \frac{1}{N}\sum_{i\in I_1^0}^N 1\left(\hat{q}_i > \gamma^0_{\max}\right) + \frac{1}{N}\sum_{i\in I_2^0}^N 1\left(\hat{q}_i < \gamma^0_{\min}\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\gamma^0_{\max} < q_i^0 + \frac{\hat{c}_i}{\sqrt{T}}\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\gamma^0_{\min} > q_i^0 + \frac{\hat{c}_i}{\sqrt{T}}\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} > \gamma^0_{\max} - q_i^0\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} < \gamma^0_{\min} - q_i^0\right).
\end{aligned}$$

Note that $\hat{c}_i > 0$ if $\hat{q}_i > q_i^0$ and $i \in I_1^0$ and $\hat{c}_i < 0$ if $\hat{q}_i < q_i^0$ and $i \in I_2^0$. Now $\gamma^0_{\max} = B_2^0$ and $\gamma^0_{\min} = B_1^0$ with $q_i^0 = B_1^0$ for those $i \in I_1^0$ and $q_i^0 = B_2^0$ for those $i \in I_2^0$,

$$\begin{aligned}
\frac{N_s}{N} &= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} > B_2^0 - q_i^0\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} < B_1^0 - q_i^0\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} > B_2^0 - B_1^0\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} < B_1^0 - B_2^0\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\hat{c}_i > \nu T^{1/2-\alpha}\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\hat{c}_i < -\nu T^{1/2-\alpha}\right)
\end{aligned}$$

Since $0 \leq \alpha < 1/2$, N_s/N tends to zero as $T \to \infty$.

II. Γ^0 and q_i^0 both unknown Now turn to the case when Γ^0 and q_i^0 are both unknown. Under the assumption that $\hat{\gamma}_{\min} - \gamma^0_{\min}$ and $\hat{\gamma}_{\max} - \gamma^0_{\max}$ are of order $O_p(N^{-1}T^{-1/2})$, we can let $\frac{\hat{d}_{\max}}{N\sqrt{T}} = \gamma^0_{\max} - \hat{\gamma}_{\max}$ and $\frac{\hat{d}_{\min}}{N\sqrt{T}} = \gamma^0_{\min} - \hat{\gamma}_{\min}$. Note that $\hat{d}_{\max}$ and $\hat{d}_{\min}$ do not depend on i. Because $\hat{q}_i = q_i^0 + \frac{\hat{c}_i}{\sqrt{T}}$, we have

$$\begin{aligned}
\frac{N_s}{N} &\leq \frac{1}{N}\sum_{i\in I_1^0}^N 1\left(\hat{q}_i > \hat{\gamma}_{\max}\right) + \frac{1}{N}\sum_{i\in I_2^0}^N 1\left(\hat{q}_i < \hat{\gamma}_{\min}\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} + \frac{\hat{d}_{\max}}{N\sqrt{T}} > \gamma^0_{\max} - q_i^0\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}} + \frac{\hat{d}_{\min}}{N\sqrt{T}} < \gamma^0_{\min} - q_i^0\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}}+\frac{\hat{d}_{\max}}{N\sqrt{T}} > B_2^0 - B_1^0\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}}+\frac{\hat{d}_{\min}}{N\sqrt{T}} < B_1^0 - B_2^0\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}}+\frac{\hat{d}_{\max}}{N\sqrt{T}} > \nu T^{-\alpha}\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\frac{\hat{c}_i}{\sqrt{T}}+\frac{\hat{d}_{\min}}{N\sqrt{T}} < -\nu T^{-\alpha}\right) \\
&= \frac{1}{N}\sum_{i\in I_1^0} 1\left(\hat{c}_i > \nu T^{-\alpha+1/2} - \frac{\hat{d}_{\max}}{N}\right) + \frac{1}{N}\sum_{i\in I_2^0} 1\left(\hat{c}_i < -\nu T^{-\alpha+1/2} - \frac{\hat{d}_{\min}}{N}\right)
\end{aligned}
$$

Notice that

$$
\begin{aligned}
P\left(\hat{q}_i > \hat{\gamma}_{\max} | i \in I_1^0\right) &= P\left(\hat{c}_i > \nu T^{-\alpha+1/2} - \frac{\hat{d}_{\max}}{N}\middle| i \in I_1^0\right) \\
&\leq P\left(|\hat{c}_i| > \nu T^{-\alpha+1/2} + O_p(N^{-1})\middle| i \in I_1^0\right) = O_p(T^{2\alpha-1}),
\end{aligned}
$$

where the last equality comes from the Chebyshev's inequality. Similarly,

$$
P\left(\hat{q}_i < \hat{\gamma}_{\min} | i \in I_2^0\right) = O_p(T^{2\alpha-1})
$$

Thus, $\mathrm{E}(N_s/N) = O_p(T^{2\alpha-1})$. Since $0 \leq \alpha < 1/2$, N_s/N tends to zero as $T \to \infty$.

Next, consider PSEUDO2 with $\hat{\gamma} = \hat{B}_\omega$. Under Assumptions 1–3, we have

$$
\begin{aligned}
P(\hat{\beta}_i > \hat{B}_\omega | \beta_i = B_1) &= P\left(\sqrt{T}(\hat{\beta}_i - B_1) > \sqrt{T}(\hat{B}_\omega - B_1)\middle|\beta_i = B_1\right) \\
&= P\left(\sqrt{T}(\hat{\beta}_i - B_1) > \sqrt{T}\left[(1-\omega)(B_2 - B_1) + O_p(\frac{1}{\sqrt{NT}})\right]\middle|\beta_i = B_1\right) \\
&= P\left((\hat{\beta}_i - B_1) > \left[(1-\omega)\nu T^{-\alpha+1/2} + O_p(\frac{1}{\sqrt{N}})\right]\middle|\beta_i = B_1\right).
\end{aligned}
$$

Also,

$$
P(\hat{\beta}_i < \hat{B}_\omega | \beta_i = B_2) = P\left((\hat{\beta}_i - B_2) > \left[\omega\nu T^{-\alpha+1/2} + O_p(\frac{1}{\sqrt{N}})\right]\middle|\beta_i = B_2\right).
$$

Similarly, under Assumptions D1–D3 with $\widetilde{x}_{it} = \widetilde{y}_{i,t-1}$, we have [10]

$$
\begin{aligned}
P(\hat{\beta}_i > \hat{B}_\omega | \beta_i = B_1) &= P\left((\hat{\beta}_i - B_1) > \left[(1-\omega)\nu T^{-\alpha+1/2} + O_p(\frac{1}{\sqrt{T}}) + O_p(\frac{1}{\sqrt{N}})\right]\middle|\beta_i = B_1\right), \\
P(\hat{\beta}_i < \hat{B}_\omega | \beta_i = B_2) &= P\left((\hat{\beta}_i - B_2) > \left[\omega\nu T^{-\alpha+1/2} + O_p(\frac{1}{\sqrt{T}}) + O_p(\frac{1}{\sqrt{N}})\right]\middle|\beta_i = B_2\right).
\end{aligned}
$$

Now $N_s = \sum_{i\in I_1^0} 1(\hat{\beta}_i > \hat{B}_\omega) + \sum_{i\in I_2^0} 1(\hat{\beta}_i < \hat{B}_\omega)$. Thus, $E(N_s/N) = P(\hat{\beta}_i > \hat{B}_\omega | \beta_i = B_1) + P(\hat{\beta}_i < \hat{B}_\omega | \beta_i = B_2) = O(T^{-1+2\alpha})$. □

[10] Under Assumptions D1-D3 with $\widetilde{x}_{it} = y_{i,t-1}$, $\hat{B}_g^0 - B_g = \sum_{i\in I_g^0} x_i' e_i / \sum_{i\in I_g^0} x_i' x_i = (1+B_g)/T + O_p(\sqrt{1/N_g^0 T})$, $g = 1, 2$, by (14), (15) and Cramer's theorem. See, for example, Alvarez and Arellano (2003, Theorem 1) for the case with $\sigma_i^2 = \sigma^2$ for all i's.

References

Abraham, C., P. Cornillion, E. Matzner-Lober, and N. Molinari (2003): "Unsupervised Curve Clustering Using B-Splines," *Scandinavian Journal of Statistics*, 30, 581–595.

Alvarez, J., and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71, 1121–1160.

Alvarez, J., M. Browning, and M. Ejrnæs (2006): "Modelling Income Processes with Lots of Heterogeneity," Oxford University Discussion Paper 285.

Anderson, T., and C. Hsiao (1982): "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics*, 18, 47–82.

Andrews, D., and W. Ploberger (1994): "Optimal Tests When a Nuisance Parameter is Present only under the Alternative," *Econometrica*, 62, 1383–1414.

Bai, J. (1997): "Estimation of a Change Point in Multiple Regression Models," *Review of Economics and Statistics*, 79, 551–563.

Bai, J., and P. Perron (1998): "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 66, 47–78.

Baltagi, B., and J. Griffin (1997): "Pooled Estimators vs. their Heterogeneous Counterparts in the Context of Dynamic Demand for Gasoline," *Journal of Econometrics*, 77, 303–327.

Baltagi, B., J. Griffin, and W. Xiong (2000): "To Pool or Not to Pool: Homogeneous versus Heterogeneous Estiatmors Applied to Cigarate Demand," *Review of Economics and Statistics*, 82, 117–126.

Barro, R. J., and X. Sala-i-Martin (2003): *Economic Growth*. The MIT Press.

Barsky, R. B., F. T. Juster, M. S. Kimball, and M. D. Shapiro (1997): "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Survey," *The Quarterly Journal of Economics*, 112, 537–579.

Browning, M., and J. Carro (2007): "Heterogeneity and Microeconometric Modeling," *Advances in Economics and Econometrics*, 3, edited by Richard Blundell, Whitney Newey and Torsten Persson, Cambridge University.

Burnside, C. (1996): "Production Function Regressions, Returns to Scale and Externalities," *Journal of Monetary Economics*, pp. 177–201.

Calinski, R., and J. J. Harabasz (1974): "A Dendrite Method for Cluster Analysis," *Communication in Statistics*, 3, 1–27.

Caner, M., and B. E. Hansen (2004): "Instrumental Variable Estimation of a Threshold Model," *Econometric Theory*, 20, 813–843.

Carroll, C. D., and A. A. Samwick (1997): "The Nature of Precautionary Wealth," *Journal of Monetary Economics*, 40, 41–71.

Chiou, J.-M., and P.-L. Li (2007): "Functional Clustering and Identifying Substructure of Longitudinal Data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 679–699.

Davies, R. B. (1977): "Hypothesis Testing when a Nuisance Parameter is Present only under the Alternative," *Biometrika*, 64, 247–254.

Duda, R., and P. Hart (1973): *Pattern Classification and Scene Analysis*. Wiley.

Durlauf, S., and P. Johnson (1995): "Multiple Regimes and Cross-Country Growth Behavior," *Journal of Applied Econometrics*, 10, 365–384.

Durlauf, S. N., A. Kourtellos, and C. M. Tan (2008): "Empirics of Growth and Development," *International Handbook of Development Economics*, 1, edited by Amitava Dutt and Jaime Ros, Edward Elgar.

Fraley, C., and A. E. Raftery (2002): "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631.

Garcia-Escudero, L., and A. Gordaliza (1999): "Robustness Properties of K Means and Trimmed K Means," *Journal of the American Statistical Association*, 94, 956–969.

Goldfeld, S., and R. Quandt (1973): "The Estimation of Structural Shifts by Switching Regressions," *Annals of Economic and Social Measurement*, 2, 475–485.

Guvenen, F. (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12, 58–79.

Hahn, J., and G. Kuersteiner (2002): "Asymptotically Unbiased Inference for A Dynamic Panel Model with Fixed Effects when Both N and T are Large," *Econometrica*, 70, 1639–1657.

Hall, P., H. Muller, and J. Wang (2006): "Properties of Principal Component Methods for Functional and Longitudinal Data Analysis," *The Annals of Statistics*, 34, 1493–1517.

Hansen, B. E. (1996): "Inference When a Nuisance Parameter Is Not Identified under the Null Hypothesis," *Econometrica*, 64, 413–430.

——— (1999): "Threshold Effects in Non-dynamic Panels: Estimation, Testing, and Inference," *Journal of Econometrics*, 93, 345–368.

Hartigan, J. A. (1975): *Clustering Algorithms*. Wiley.

Henderson, D. J., and R. R. Russell (2005): "Human Capital and Convergence: A Production-Frontier Approach," *International Economic Review*, 46, 1167–1205.

Heston, A., R. Summers, and B. Aten (2006): *Penn World Table Version 6.2*. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Hsiao, C., and M. H. Pesaran (2004): "Random Coefficient Panel Data Models," edited by L. Matyas and P. Sevestre, Third Edition, Springer Publishers, Ch. 6.

Hsiao, C., and A. K. Tahmiscioglu (1997): "A Panel Analysis of Liquidity Constraints and Firm Investment," *Journal of the American Statistical Association*, 92, 455–465.

Juárez, M. A., and M. F. J. Steel (2010): "Model-based Clustering of non-Gaussian Panel Data Based on Skew-t Distributions," *Journal of Business and Economic Statistics*, 28, 52–66.

Lawrance, E. C. (1991): "Poverty and the Rate of Time Preference: Evidence from Panel Data," *The Journal of Political Economy*, 99, 54–77.

Lee, K., M. H. Pesaran, and R. Smith (1997): "Growth and Convergence in a Multi-country Empirical Stochastic Solow Model," *Journal of Applied Econometrics*, 12, 357–392.

Maddala, G., R. Trost, H. Li, and F. Joutz (1997): "Estimationn of Short Run and Long Run Elasticities of Energy Demand from Panel Data using Shrinkage Estimators," *Journal of Business and Economic Statistics*, 15, 90–100.

Mankiw, N. G., D. Romer, and D. N. Weil (1992): "A Contribution to the Emprics of Economic Growth," *The Quarterly Journal of Economics*, 107, 407–437.

Milligan, G., and W. Cooper (1985): "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, 50, 159–179.

Pesaran, M. H. (2006): "Estimation and inference in large heterogeneous panels with a multi-factor error structure," *Econometrica*, 74(4), 967–1012.

Pesaran, M. H., and T. Yamagata (2008): "Testing Slope Homogeneity in Large Panels," *Journal of Econometrics*, 142, 50–93.

Pollard, D. (1981): "Strong Consistency of K-means Clustering," *Annals of Statistics*, 9, 135–140.

——— (1982): "A Central Limit Theorem for K-Means Clustering," *Annals of Probability*, 10, 919–926.

Robertson, D., and J. Symons (1992): "Some Strange Properties of Pooled Data Estimators," *Journal of Applied Econometrics*, 7, 175–189.

Sugar, C., and G. James (2003): "Finding the Number of Clsuters in a Dataset: An Information Theoretic Approach," *Journal of the American Statistical Association*, 98:463, 750–763.

Sun, Y. (2005): "Estimation and Inference in Panel Structure Models," *Working Paper:* 2005-11, Dept. of Economics, University of California, San Diego.

Swamy, P. A. V. B. (1970): "Efficient Inference in a Random Coefficient Regression Model," *Econometrica*, 38, 311–323.

Figure 1: RMSEs of the proposed methods

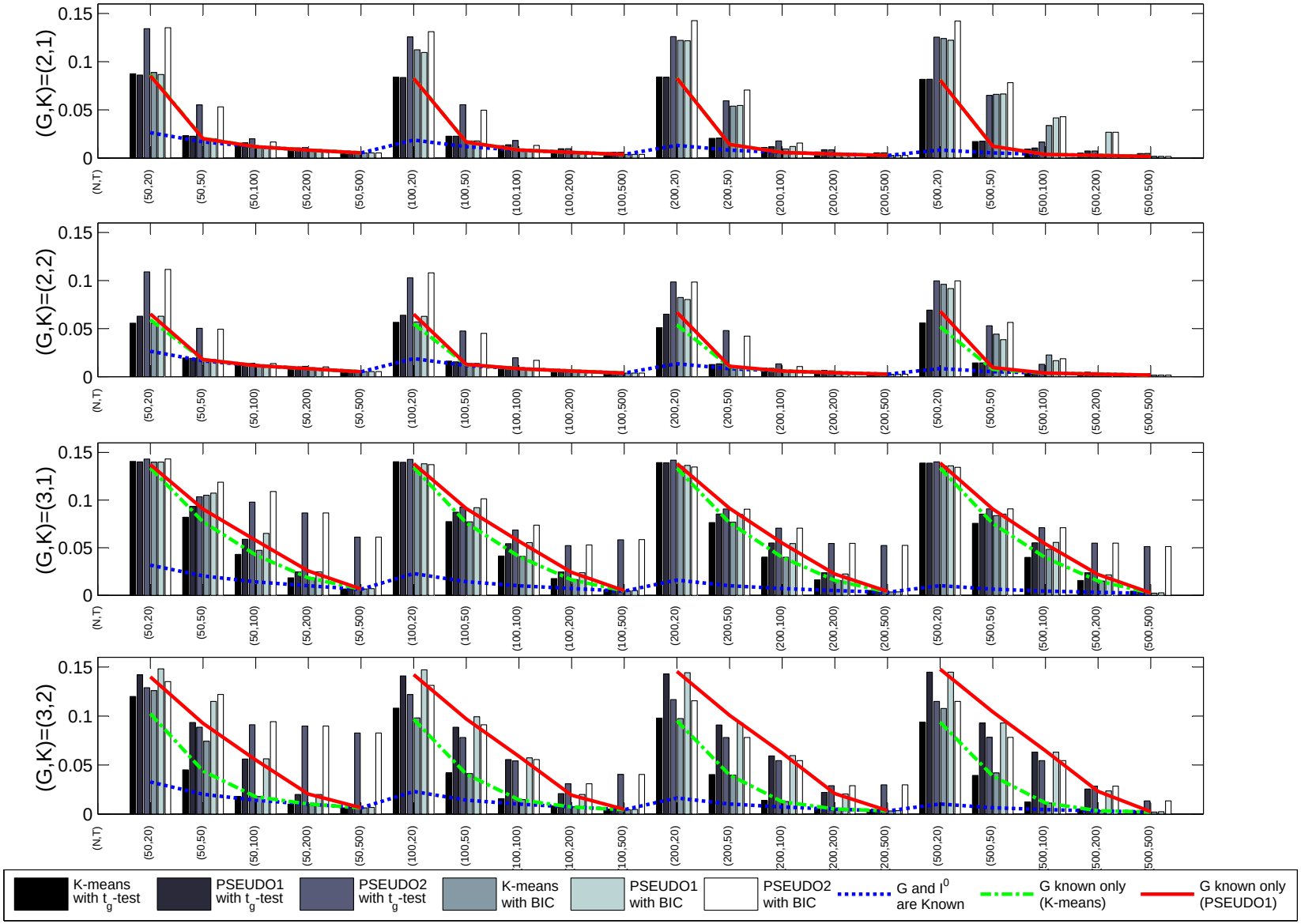

Figure 2: Probability of Selecting the Correct Group Number

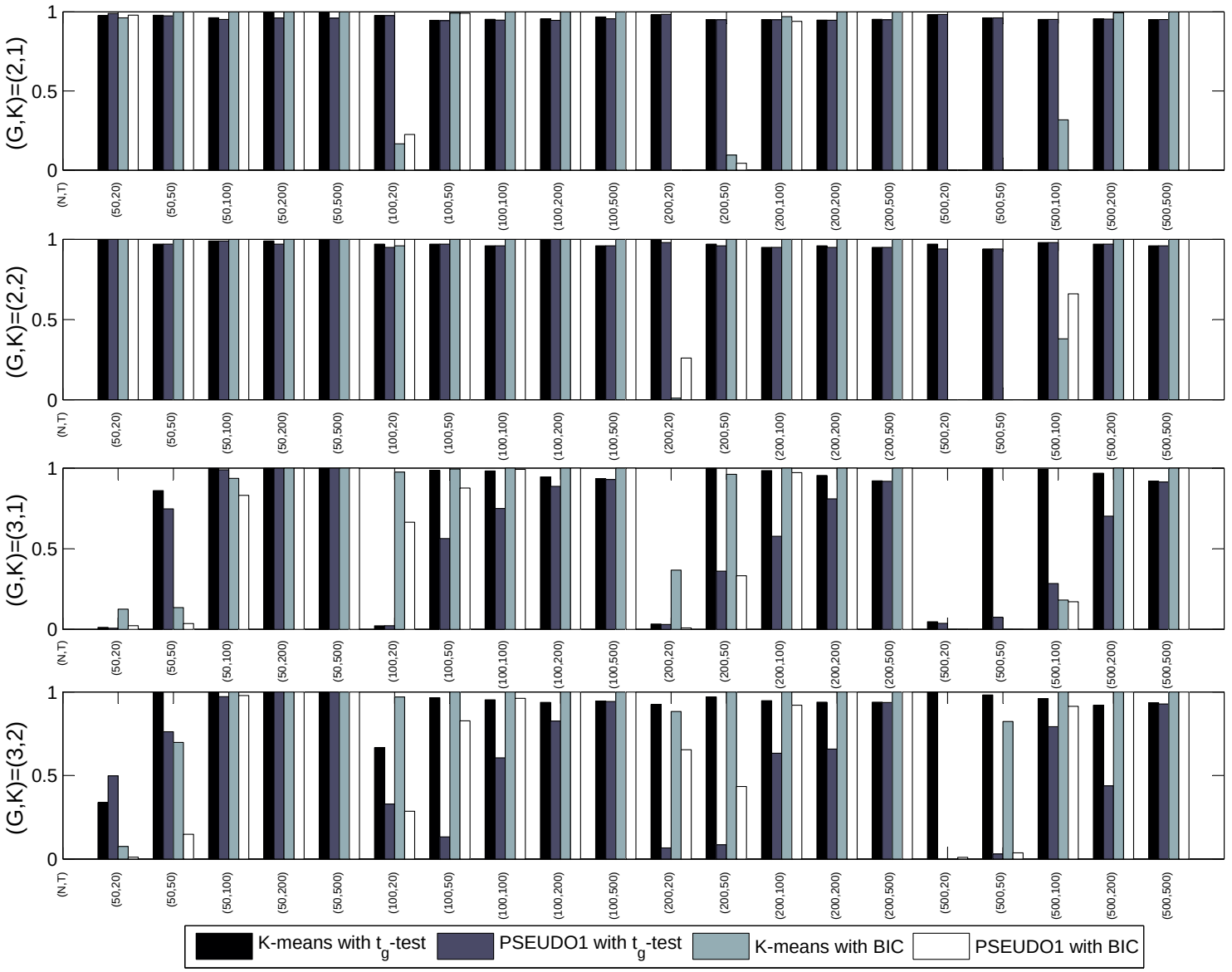

Figure 3: Accuracy of Assigning Group Membership

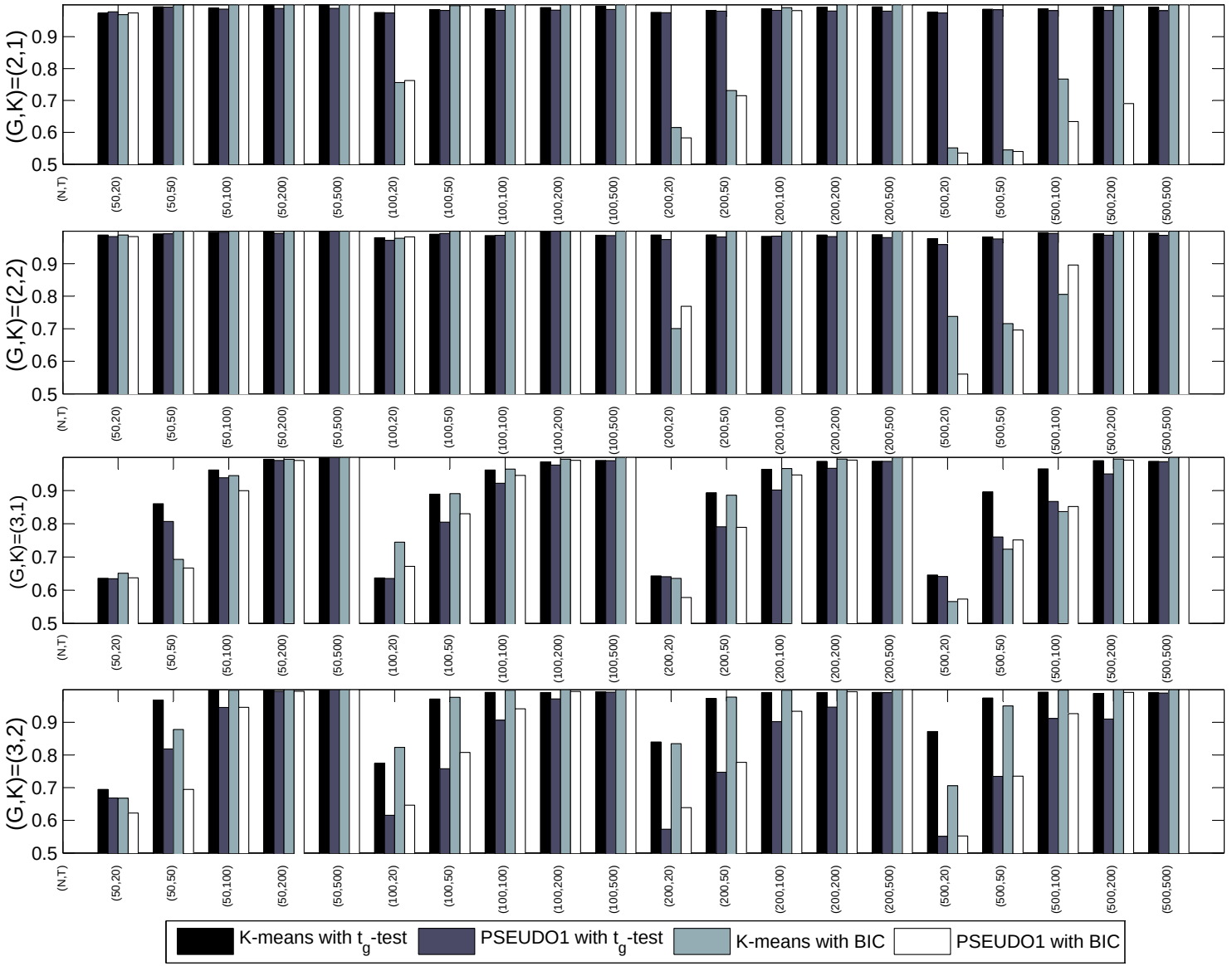

Figure 4: Out of Sample Sum of Squared Residuals

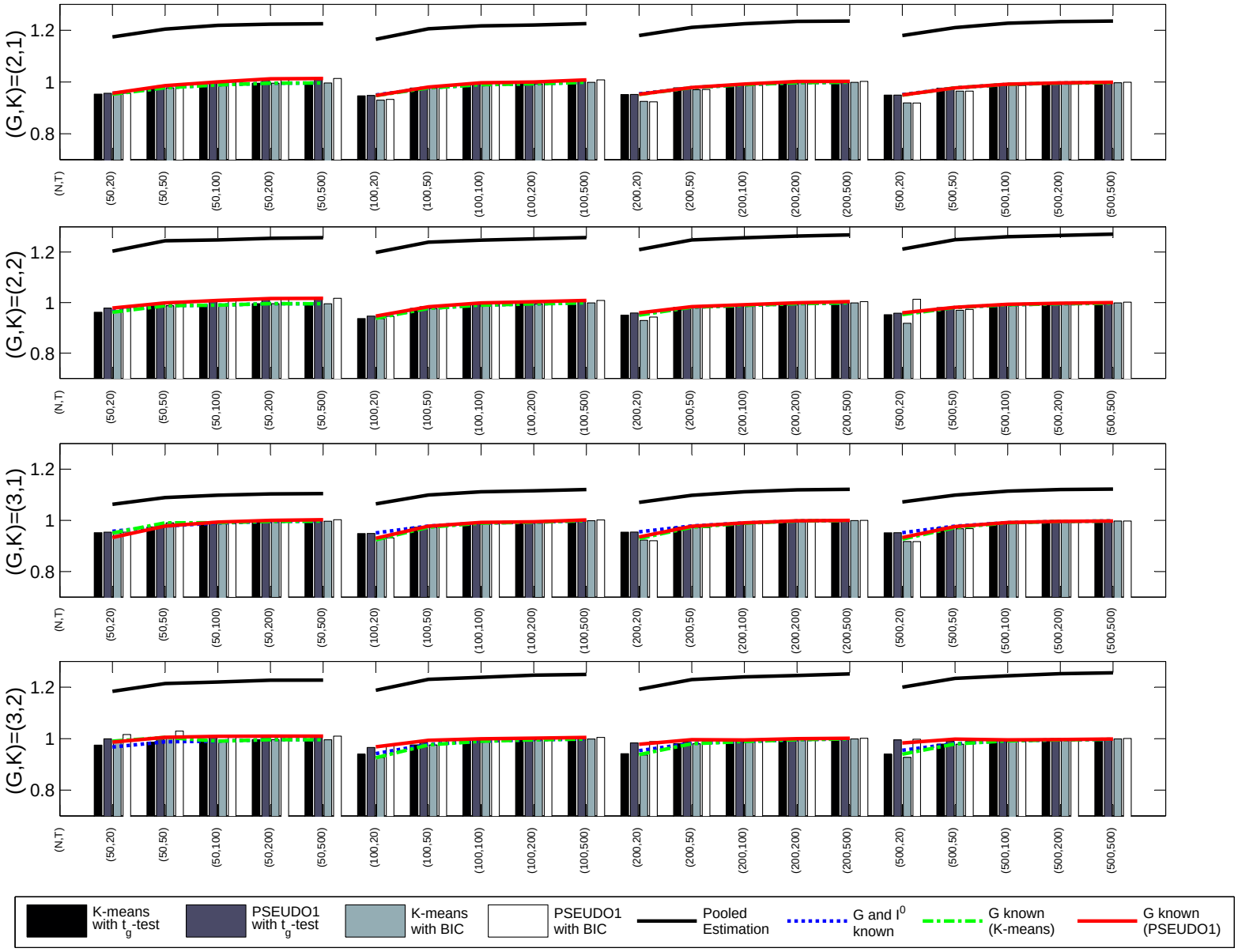

Table 1: Results for Dynamic Panel Model.

t_g-test										
RMSE×10										
	K-means					PSEUDO1				
	T=20	50	100	200	500	T=20	50	100	200	500
N=50	1.35	0.35	0.19	0.12	0.07	1.55	0.62	0.18	0.12	0.07
100	1.36	0.41	0.22	0.12	0.06	1.59	0.73	0.22	0.11	0.05
200	1.40	0.46	0.25	0.12	0.06	1.63	0.79	0.23	0.12	0.06
500	1.50	0.50	0.31	0.15	0.06	1.65	0.72	0.29	0.13	0.05
Probability of Selecting the Correct Group Number (GR)										
	K-means					PSEUDO1				
	T=20	50	100	200	500	T=20	50	100	200	500
N=50	0.88	0.89	0.91	0.94	0.95	0.79	0.82	0.92	0.95	0.95
100	0.73	0.66	0.83	0.91	0.96	0.43	0.62	0.83	0.91	0.96
200	0.45	0.42	0.72	0.87	0.93	0.20	0.65	0.72	0.87	0.93
500	0.07	0.08	0.44	0.78	0.93	0.12	0.37	0.44	0.78	0.93
Accuracy of Classification(CR)										
	K-means					PSEUDO1				
	T=20	50	100	200	500	T=20	50	100	200	500
N=50	0.85	0.92	0.94	0.96	0.97	0.75	0.86	0.94	0.97	0.96
100	0.78	0.80	0.88	0.93	0.97	0.62	0.74	0.88	0.94	0.97
200	0.63	0.64	0.79	0.90	0.95	0.51	0.75	0.80	0.91	0.96
500	0.39	0.44	0.59	0.83	0.95	0.46	0.61	0.59	0.84	0.95
BIC										
RMSE×10										
	K-means					PSEUDO1				
	T=20	50	100	200	500	T=20	50	100	200	500
N=50	1.31	0.30	0.16	0.10	0.06	1.54	0.58	0.16	0.10	0.06
100	1.34	0.27	0.14	0.08	0.04	1.60	0.67	0.14	0.08	0.04
200	1.55	0.25	0.12	0.07	0.03	1.66	0.76	0.13	0.07	0.03
500	1.59	0.52	0.28	0.06	0.03	1.65	0.72	0.23	0.06	0.03
Probability of Selecting the Correct Group Number (GR)										
	K-means					PSEUDO1				
	T=20	50	100	200	500	T=20	50	100	200	500
N=50	0.98	1.00	1.00	1.00	1.00	0.91	1.00	1.00	1.00	1.00
100	0.80	1.00	1.00	1.00	1.00	0.40	0.96	1.00	1.00	1.00
200	0.00	1.00	1.00	1.00	1.00	0.07	0.89	1.00	1.00	1.00
500	0.00	0.00	0.20	1.00	1.00	0.10	0.37	0.26	1.00	1.00
Accuracy of Classification(CR)										
	K-means					PSEUDO1				
	T=20	50	100	200	500	T=20	50	100	200	500
N=50	0.89	1.00	1.00	1.00	1.00	0.80	0.98	1.00	1.00	1.00
100	0.80	1.00	1.00	1.00	1.00	0.62	0.95	1.00	1.00	1.00
200	0.35	1.00	1.00	1.00	1.00	0.43	0.91	1.00	1.00	1.00
500	0.31	0.31	0.46	1.00	1.00	0.44	0.60	0.50	1.00	1.00

Table 2: Application: Growth Regression

Model A								
Group	1		2		3		4	
	K-means	PSUEDO1	K-means	PSUEDO1	K-means	PSUEDO1	K-means	PSUEDO1
N_g	9	9	20	13	27	23	13	24
ρ_g	0.8394	0.8663	0.9727	0.9693	0.8764	0.9178	0.9519	0.9592
	(31.052)	(35.858)	(166.939)	(125.573)	(73.124)	(79.362)	(93.948)	(123.846)
ϕ_g	-0.0009	-0.0009	-0.0004	-0.0005	0.0018	0.0011	0.0018	0.0009
	(-3.074)	(-3.423)	(-2.577)	(-2.318)	(7.369)	(4.412)	(4.893)	(3.858)
t_g-test	0.5454	2.9625	1.2584	-0.2241	1.1865	0.9209	0.0757	6.3712
Model B								
Group	1		2		3		4	
N_g	8	9	11	20	14	19	36	21
ρ_g	0.8771	0.8736	0.9279	0.9606	0.924	0.8866	0.9791	0.9681
	(42.684)	(41.152)	(57.232)	(114.852)	(73.169)	(59.59)	(201.029)	(131.614)
ϕ_g	0.0009	0.0011	-0.0004	-0.0002	0.0021	0.0026	0.0004	0.0011
	(2.133)	(3.130)	(-1.477)	(-1.081)	(6.244)	(7.375)	(3.263)	(4.695)
t_g-test	0.1194	-0.3125	-0.0374	2.8038	6.0253	5.1246	1.6078	3.2935

Note: Regression models for $g = 1, 2, \ldots, G$:

$$\begin{aligned} \text{Model A}: \ \widetilde{y}_{it} &= \alpha_i + \rho_g \widetilde{y}_{it-1} + \phi_g t + \widetilde{e}_{it}, \quad i \in I_g, \\ \text{Model B}: \ \widetilde{y}_{it} &= \alpha_i + \rho_g \widetilde{y}_{it-1} + \phi_g t + \beta_g \bar{\triangle} y_{t-1} + \gamma_g \bar{\triangle} y_{t-2} + \widetilde{e}_{it}, \quad i \in I_g, \end{aligned}$$

where α_i denotes country-specific fixed effects, ρ_g, ϕ_g, β_g and γ_g, $g = 1, \ldots, 4$, are group specific, $\bar{y}_s = N^{-1} \sum_i^N \widetilde{y}_{is}$, $\bar{\triangle} y_s = \bar{y}_s - \bar{y}_{s-1}$, $s = 1, \ldots, T$. The number in parentheses is the t-value. The detailed definition of the t_g-test can be found in Equation (9).