To Pool or Not to Pool: A Partially Heterogeneous Framework

Sarafidis, Vasilis and Weber, Neville

The University of Sydney

8 December 2009

To Pool or Not to Pool: A Partially Heterogeneous Framework[*]

Vasilis Sarafidis[†]
University of Sydney

Neville Weber[‡]
University of Sydney

This version: December 2009

Abstract

This paper proposes a partially heterogeneous framework for the analysis of panel data with fixed T, based on the concept of 'partitional clustering'. In particular, the population of cross-sectional units is grouped into clusters, such that parameter homogeneity is maintained only within clusters. To determine the (unknown) number of clusters we propose an information-based criterion, which, as we show, is strongly consistent $-$ i.e. it selects the true number of clusters with probability one as $N \to \infty$. Simulation experiments show that the proposed criterion performs well even with moderate N and the resulting parameter estimates are close to the true values. We apply the method in a panel data set of commercial banks in the US and we find four clusters, with significant differences in the slope parameters across clusters.

Key Words: partial heterogeneity, partitional clustering, information-based criterion, model selection.

JEL Classification: C13; C33; C51.

[*] We would like to thank Geert Dhaene, Tom Wansbeek, and seminar participants at the Tinbergen Institute, University of York, University of Leuven and Erasmus University Rotterdam for helpful comments and suggestions. We gratefully acknowledge financial support from the Research Unit of the Faculty of Economics and Business, University of Sydney.

[†] Corresponding author. Faculty of Economics and Business, University of Sydney, NSW 2006, Australia. Tel: +61-2-9036 9120; E-mail: v.sarafidis@econ.usyd.edu.au.

[‡] School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia. Tel: +61-2-9351 4249; E-mail: n.weber@maths.usyd.edu.au.

1 Introduction

Full homogeneity in the slope coefficients of a panel data model is often an assumption that is difficult to justify, both on theoretical grounds and from a practical point of view. On the other hand, the alternative of imposing no structure on how these coefficients may vary across individuals may be rather extreme. This argument is in line with evidence provided by a substantial body of applied work. For example, Baltagi and Griffin (1997) reject the hypothesis of coefficient homogeneity in a panel of gasoline demand regressions across the OECD countries, and Burnside (1996) rejects the hypothesis of homogeneous production function parameters in a panel of US manufacturing industries. Even so, both studies show that fully heterogeneous models lead to very imprecise estimates of the parameters, which in some cases have even the wrong sign. Baltagi and Griffin notice that this is the case despite the fact that there is a relatively long time series in the panel – to the extent that the traditional pooled estimators are superior in terms of root mean square error and forecasting performance. Furthermore, Burnside suggests that in general the results of his estimates show significant differences between the fully homogeneous and the fully heterogeneous models and the conclusions about the degree of returns to scale in the manufacturing industry would heavily depend on which one of these two models is used. In the same line Baltagi, Griffin and Xiong (2000) place the debate between homogeneous versus heterogeneous panel estimators in the context of cigarette demand and conclude that even with T (the number of time series observations) relatively large, heterogeneous models for individual states tend to produce implausible estimates with inferior forecasting properties, despite the fact that parameter homogeneity is soundly rejected by the data. Similar conclusions are reached by Baltagi, Bresson and Pirotte (2002) using evidence from US electricity and gas consumption.

These findings indicate that the modelling framework of complete homogeneity (pooling) and full heterogeneity may be polar cases, and other intermediate cases may often provide more realistic solutions in practice. The pooled mean group estimator (PMGE) proposed by Pesaran, Shin and Smith (1999) is a formal attempt to bridge the gap between pooled and fully heterogeneous estimators by imposing partially heterogeneous restrictions related to the time dimension of the panel. In particular, this intermediate estimator allows the short-run parameters of the model to be individual-specific and restricts the long-run coefficients to be the same across individuals for reasons attributed to budget constraints, arbitrage conditions and common technologies. This procedure is appealling because it imposes constraints that can be directly related to economic theory, although it is mainly designed for panels where both dimensions are large.

In this paper we put forward a modelling framework that imposes partially heterogeneous restrictions not with respect to the time dimension of the panel, as PMGE does, but with respect to the cross-sectional dimension, N. In particular, the population of cross-sectional units is grouped into distinct clusters, such that within each cluster the parameters are homogeneous and all intra-cluster heterogeneity is attributed to the usual individual-specific unobserved effects. The clusters themselves are heterogeneous – that is, the slope coefficients vary across clusters.

Naturally, the practical issue of how to cluster the individuals into homogeneous groups is central in the paper. Clustering methods have already been advocated in the econometric panel data literature by some researchers; for instance, Durlauf and Johnson (1995) propose clustering the individuals using regression tree analysis, and Vahid (1999) suggests a classification algorithm based on a measure of complexity using the principles of minimum description length and minimum message length, often employed in coding theory.[1] Both these methods are based on the concept of *hierarchical clustering*, which involves building a 'hierarchy' from the individual units by progressively merging them into larger clusters. As a result, the proposed algorithms are theoretically founded for $T \to \infty$ only. On the other hand, when T is small these algorithms can have poor properties in terms of determining the appropriate number of clusters. Thus, Vahid (1999) concludes:

> "The classic homogeneity assumption in panel data analysis ... is absolutely necessary and non-testable for the analysis of panel data with very small T", page 413.

On the contrary, this paper proposes estimating the unknown number of clusters, as well as the corresponding partition, using an information-based criterion that is consistent for any fixed T – that is, the probability of estimating the true number of clusters approaches one as $N \to \infty$. This is important because most

[1] Recently, Kapetanios (2006) proposed an information criterion, based on simulated annealing, to address a related problem – in particular, how to decompose a set of series into a set of poolable series for which there is evidence of a common parameter subvector and a set of series for which there is no such evidence.

frequently panel data sets entail a large number of individuals and a small number of time series observations. Furthermore, it is usually the case when T is small where some kind of pooling provides substantial efficiency gains over full heterogeneity. Our method relies on the concept of *partitional clustering*; instead of treating each individual as a distinct cluster to begin with (as in hierarchical clustering), the underlying structure is recovered from the data by grouping the individuals into a fixed number of clusters using an initial partition, and then re-allocating each individual into the remaining clusters such that the final preferred partition minimises an objective function. In this paper the residual sum of squares (RSS) of the estimated model is used as the objective function. The number of clusters is determined by the clustering solution that minimises RSS subject to a penalty function that is strictly increasing in the number of clusters. The penalty reflects the fact that the minimum RSS of the estimated model is monotone decreasing in the number of clusters and therefore it tends to over-parameterise the model by allowing for more clusters than there actually exist. Hence, the penalty acts essentially as a filter to ensure that the preferred clustering outcome partitions between clusters rather than within clusters. The intuition of the procedure is identical to a standard model selection criterion, although the study of the asymptotics is more complicated in our case because the number of individuals contained in a given cluster may vary with N.

The remainder of the paper is as follows. The next section sets out our partially heterogeneous model and distinguishes the concept of clustering proposed in the paper from other clustering concepts known in the literature. Section 3 examines the properties of the pooled fixed effects and OLS estimators under partial heterogeneity. Section 4 formulates the clustering problem, analyses the objective function and discusses the proposed partitional clustering algorithm. The finite-sample performance of the algorithm is investigated in Section 5 using Monte Carlo experiments. Section 6 illustrates the technique using a random panel of 551 banking institutions operating in the US, each observed over a period of 15 years. A final section concludes.

2 Model Specification

We consider the following panel data model:

$$y_{\omega it} = \boldsymbol{\beta}'_{\omega}\mathbf{x}_{\omega it} + \kappa_{\omega} + \eta_{\omega i} + \varepsilon_{\omega it}, \tag{1}$$

where $y_{\omega it}$ denotes the observation on the dependent variable for the i^{th} individual that belongs to cluster ω at time t, $\boldsymbol{\beta}_{\omega} = (\beta_{\omega 1}, ..., \beta_{\omega K})'$ is a $K \times 1$ vector of fixed unknown coefficients, $\mathbf{x}_{\omega it} = (x_{\omega it1}, ..., x_{\omega itK})'$ is a $K \times 1$ vector of covariates, κ_{ω} and $\eta_{\omega i}$ denote cluster- and individual-specific time-invariant effects respectively and $\varepsilon_{\omega it}$ is a purely idiosyncratic error component. We have $\omega = 1, ..., \Omega$, $i\,[\in \omega] = 1, ..., N_{\omega}$, and $t = 1, ..., T$. This means that the total number of clusters equals Ω, the ω^{th} cluster has N_{ω} individuals, for which there are T time series observations available. The total number of individuals in all clusters equals $N = \sum_{\omega=1}^{\Omega} N_{\omega}$ and the total sample size is given by $S = NT$.

Essentially, model (1) makes a case for pooling the data only within clusters of individuals and allowing for slope heterogeneity across clusters. One way to rationalise this is on the basis of the existence of certain unobserved factors or qualities, which decompose the population of individuals into distinct groups, such that within each group individuals respond similarly to changes in the regressors and all intra-cluster heterogeneity is attributed to individual-specific time-invariant effects, while individuals from different clusters may differ in terms of these factors/qualities and therefore they behave in a different manner. There are several examples where this set up may apply in practice. For instance, in a model of economic growth it is natural to think that growth determinants have different marginal impacts for different groups of countries although the number and size of the clusters is typically unknown. Indeed, recent theories of growth and development (e.g. Galor, 1996, and Temple, 1999) suggest the presence of convergence clubs without specifying club membership or the number of clubs. Another example can be drawn from the financial industry, where existing evidence (see e.g. Berger and Humphrey, 1997) appear to suggest that the underlying technology varies across banking institutions of different size and therefore pooling the data may lead to misleading inferences about the returns to scale in the industry.

It is useful to distinguish the concept of clustering proposed in this paper from other clustering methods or structures of data already familiar in the econometric literature. In particular, it is common to assume that the errors are independent across clusters but not within clusters, which gives rise to 'clustered standard errors'. Neglecting this feature in the data and using OLS-type estimates of the standard errors is likely to lead to biased inferences, although the first-order properties of the estimator remain unaffected[2]. There is also some resemblance between our partially heterogeneous model (1) and 'hierarchical' or 'nested' or

[2] See e.g. Cameron and Trivedi (2005, Section 24.5).

'multi-level' structures, which allow for a hierarchy of the observations among different levels of data (see e.g. Antweiler, 2001, and Baltagi, Song and Jung, 2001). For example, $y_{\omega it}$ could denote the observation of air polution measured by station i, which is located in city ω, at time t and so on. Our partially heterogeneous model can then be viewed as a single-nested structure. The main difference is that in a multi-level structure the number of clusters and the corresponding partition are known and the emphasis is on estimating a multi-way error components model[3], while in our framework the total number of clusters, as well as individual membership into these clusters, need to be estimated. Our partially heterogeneous model is also conceptually similar to a mixture regression model, except that the latter involves devising an appropriate probabilistic mechanism that determines membership of individual i into cluster ω[4], and requires specifying distributional assumptions for the components of the mixture. Our model selection criterion is relatively parsimonious and distribution-free.

We make the following Basic Assumptions (**BA**):

BA.1 $\varepsilon_{\omega it}$ is uncorrelated across ω and i, with $E\left(\varepsilon_{\omega it}|\mathbf{x}_{\omega i1},...,\mathbf{x}_{\omega iT}\right) = 0$ and $E\left(\varepsilon^2_{\omega it}|\mathbf{x}_{\omega i1},...,\mathbf{x}_{\omega iT}\right) = \sigma^2_{\varepsilon_\omega} < \infty \; \forall \; \omega$ and i.

BA.2 $N_\omega^{-1}\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i} \xrightarrow{p} M_{XX,\omega}$, finite and positive definite, where $X_{\omega i} = \left(\mathbf{x}_{\omega i1},...,\mathbf{x}_{\omega iT}\right)'$.

BA.3 There exists a fixed constant, $0 < c_\omega < 1$, such that $N_\omega/N \to c_\omega$ for $\omega = 1,...,\Omega$, as $N \to \infty$.

BA.4 Ω is a fixed unknown integer, such that $0 < \Omega \leq \xi$, where ξ is fixed and known.

Assumptions BA.1 and BA.2 are standard in the analysis of panel data models with strongly exogenous regressors. For example, BA.2 ensures that $\left(N_\omega^{-1}\sum_{i=1}^{N_\omega} X'_{\omega i}X_{\omega i}\right)^{-1}$ exists in probability for all N_ω sufficiently large. BA.3 ensures that no clusters are asymptotically negligible. The asymptotics can be conceived via 'class-growing sequences', as in Shao and Wu (2005). Assumption BA.4 ensures that the total number of clusters is bounded by a known integer, ξ. Observe that we have

[3] Hsiao (2003, Section 10.4) and Baltagi (2008, Section 8.6) provide a concise introduction of multi-level structures.

[4] For example, for $K = 1$ and $\Omega = 2$, a mixture regression model will estimate 4 parameters, two cluster-specific slope coefficients together with the associated mixture weights.

not imposed any restrictions regarding the distribution of $\eta_{\omega i}$ (and κ_{ω}), which are allowed to be correlated with $\mathbf{X}_{\omega i}$, or the serial correlation properties of $\varepsilon_{\omega it}$.

We define

$$\boldsymbol{\beta}_{\omega} = \boldsymbol{\beta} + \boldsymbol{\delta}_{\omega}, \tag{2}$$

where $\boldsymbol{\delta}_{\omega}$ is a $K \times 1$ vector of fixed constants, such that $\sum_{\omega=1}^{\Omega} c_{\omega}\boldsymbol{\delta}_{\omega} = \mathbf{0}$. (2) implies that $\boldsymbol{\beta}$ is a weighted average of the cluster-specific coefficients, $\boldsymbol{\beta}_{\omega}$, with the weights depending on the proportion of individuals that each cluster contains in the "long term", i.e. as N grows.

Without any information upon (i) cluster membership and (ii) the size of Ω, one can only obtain an estimate of $\boldsymbol{\beta}$, and the next section addresses whether $\boldsymbol{\beta}$ can be estimated consistently using the standard fixed effects and pooled OLS estimators.

3 On the Impact of Partial Heterogeneity

Model (1) can be expressed in vector form as follows:

$$\mathbf{y}_{\omega i} = X_{\omega i}\boldsymbol{\beta}_{\omega} + \iota_T \eta_{\omega i} + \boldsymbol{\varepsilon}_{\omega i}, \tag{3}$$

where $\mathbf{y}_{\omega i} = (y_{\omega i1}, ..., y_{\omega iT})'$, $X_{\omega i} = (\mathbf{x}_{\omega i1}, ..., \mathbf{x}_{\omega iT})'$, $\boldsymbol{\varepsilon}_{\omega i} = (\varepsilon_{\omega i1}, ..., \varepsilon_{\omega iT})'$, ι_T is a $T \times 1$ vector of ones and without loss of generality we have imposed $\kappa_{\omega} = 0$.[5]

Ignoring the partially heterogeneous structure in (3) results in the following regression model:

$$\mathbf{y}_{\omega i} = X_{\omega i}\boldsymbol{\beta} + \iota_T \eta_{\omega i} + \boldsymbol{\upsilon}_{\omega i},\ \boldsymbol{\upsilon}_{\omega i} = \boldsymbol{\varepsilon}_{\omega i} + X_{\omega i}\boldsymbol{\delta}_{\omega}. \tag{4}$$

Define the $T \times T$ idempotent matrix $Q_T = I_T - T^{-1}\boldsymbol{\iota}_T\boldsymbol{\iota}_T'$, which transforms the observations in terms of deviations from individual-specific averages and sweeps out the time-invariant effects, $\eta_{\omega i}$. We have

$$Q_T\mathbf{y}_{\omega i} = Q_T X_{\omega i}\boldsymbol{\beta} + Q_T\boldsymbol{\upsilon}_{\omega i},\ Q_T\boldsymbol{\upsilon}_{\omega i} = Q_T\boldsymbol{\varepsilon}_{\omega i} + Q_T X_{\omega i}\boldsymbol{\delta}_{\omega}, \tag{5}$$

or

$$\widetilde{\mathbf{y}}_{\omega i} = \widetilde{X}_{\omega i}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}_{\omega i} + \widetilde{X}_{\omega i}\boldsymbol{\delta}_{\omega}, \tag{6}$$

where $\widetilde{\mathbf{y}}_{\omega i} = Q_T\mathbf{y}_{\omega i}$ and similarly for the remaining variables.

[5] This is not a restriction because one can always define $\eta_{\omega i}^{*} = \eta_{\omega i} + k_{\omega}$.

The fixed effects estimator is given by

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{FE} &= \left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{X}_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{\mathbf{y}}_{\omega i}\right] \\ &= \boldsymbol{\beta}+\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{X}_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\left(\widetilde{X}'_{\omega i}\widetilde{\boldsymbol{\varepsilon}}_{\omega i}+\widetilde{X}'_{\omega i}\widetilde{X}_{\omega i}\boldsymbol{\delta}_\omega\right)\right].\end{aligned} \tag{7}$$

Taking plims over N yields

$$\begin{aligned}&\text{plim}_{N\to\infty}\left(\widehat{\boldsymbol{\beta}}_{FE}-\boldsymbol{\beta}\right)= \\ &= \text{plim}_{N\to\infty}\left[\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{X}_{\omega i}\right)\right]^{-1}\left\{\text{plim}_{N\to\infty}\left[\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{\boldsymbol{\varepsilon}}_{\omega i}\right)\right]\right. \\ &\quad\left.+\ \text{plim}_{N\to\infty}\left[\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{X}_{\omega i}\right)\boldsymbol{\delta}_\omega\right]\right\} \\ &= \left[\sum_{\omega=1}^{\Omega}c_\omega\widetilde{M}_{XX,\omega}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\widetilde{M}_{XX,\omega}c_\omega\boldsymbol{\delta}_\omega\right],\end{aligned} \tag{8}$$

where $\widetilde{M}_{XX,\omega}=\text{plim}_{N_\omega\to\infty}\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{X}'_{\omega i}\widetilde{X}_{\omega i}$, the existence of which is guaranteed from BA.2. As we can see from (8), the fixed effects estimator is not necessarily consistent. In particular, consistency is achieved when $\widetilde{M}_{XX,\omega}c_\omega\boldsymbol{\delta}_\omega$ sums up to zero and this will occur, for example, when the limiting matrix $\widetilde{M}_{XX,\omega}$ is orthogonal to the vector $c_\omega\boldsymbol{\delta}_\omega$ for all ω. This would be the case if, say, the $\widetilde{M}_{XX,\omega}$ matrices are constant across clusters. However, this condition is unnatural in economic data sets and therefore it is unlikely to hold true in most empirical applications.

The above result may be quite surprising because in a fully heterogeneous model, where each individual forms its own cluster such that the model becomes[6]

$$y_{it}=\alpha_i+\boldsymbol{\beta}_i'\mathbf{x}_{it}+\varepsilon_{it}\text{ with }\boldsymbol{\beta}_i=\boldsymbol{\beta}+\boldsymbol{\delta}_i, \tag{9}$$

exogeneity of the regressors, namely $E\left(\boldsymbol{\delta}_i|\mathbf{x}_{i1},...,\mathbf{x}_{iT}\right)=0$, is sufficient to ensure that $\widehat{\boldsymbol{\beta}}_{FE}$ is consistent, although not efficient. Similarly, in a fully homogeneous model consistency follows because $\boldsymbol{\delta}_\omega=0$. However, as we see here, in the intermediate case of partial heterogeneity $\widehat{\boldsymbol{\beta}}_{FE}$ does not converge to $\boldsymbol{\beta}$ in general. In other words, pooling the data when the slope parameters are partially heterogeneous has first- and second-order implications for the fixed effects estimator. When T is sufficiently large, one may deal with this problem by estimating

[6] The subscript ω is omitted in this case because $\Omega=N$.

individual-specific regressions and clustering the individuals (if required) on the basis of their estimated coefficients. However, for T fixed this approach may not even be feasible.

A similar result holds for the pooled OLS estimator, although in this case the properties of the estimator also depend on the mean value of $\mathbf{x}_{\omega it}$. To illustrate the main idea, it is convenient to assume that the $\eta_{\omega i}$ term is uncorrelated and $E\left(\eta_{\omega it}|\mathbf{x}_{\omega i1},...,\mathbf{x}_{\omega iT}\right)=0\ \forall\ \omega$ and i.

The estimable model is now given by

$$y_{\omega it}=\boldsymbol{\beta}'\mathbf{x}_{\omega it}+\left(\varepsilon_{\omega it}+\eta_{\omega i}+\boldsymbol{\delta}'_{\omega}\mathbf{x}_{\omega it}\right), \tag{10}$$

or, in matrix form,

$$\mathbf{y}_{\omega i}=X_{\omega i}\boldsymbol{\beta}+\mathbf{v}_{\omega i},\ \mathbf{v}_{\omega i}=\boldsymbol{\varepsilon}_{\omega i}+\iota_{T}\eta_{\omega i}+X_{\omega i}\boldsymbol{\delta}_{\omega}.$$

Hence, the pooled OLS estimator of $\boldsymbol{\beta}$ is

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= \left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_{\omega}}X'_{\omega i}X_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_{\omega}}X'_{\omega i}\mathbf{y}_{\omega i}\right]\\ &= \boldsymbol{\beta}+\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_{\omega}}X'_{\omega i}X_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_{\omega}}X'_{\omega i}\left(\boldsymbol{\varepsilon}_{\omega i}+\iota_{T}\eta_{\omega i}+X_{\omega i}\boldsymbol{\delta}_{\omega}\right)\right]. \end{aligned}\tag{11}$$

Taking plims over N yields[7]

$$\begin{aligned}&\text{plim}_{N\to\infty}\left(\widehat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}\right)=\\ &=\left[\sum_{\omega=1}^{\Omega}\left(\widetilde{M}_{XX,\omega}+T\widetilde{W}_{XX,\omega}\right)c_{\omega}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\left(\widetilde{M}_{XX,\omega}+T\widetilde{W}_{XX,\omega}\right)c_{\omega}\boldsymbol{\delta}_{\omega}\right],\end{aligned}\tag{12}$$

where $\widetilde{W}_{XX,\omega}=\text{plim}_{N_{\omega}\to\infty}\frac{1}{N_{\omega}}\sum_{i=1}^{N_{\omega}}\overline{X}'_{\omega i}\overline{X}_{\omega i}$ with $\overline{X}_{\omega i}=T^{-1}\boldsymbol{\iota}'_{T}X_{\omega i}$, while $\widetilde{M}_{XX,\omega}$ has been already defined below (8). Hence, comparing (8) with (12) we see that the difference between the two estimators is that while a sufficient condition for consistency of $\widehat{\boldsymbol{\beta}}_{FE}$ is that $\widetilde{M}_{XX,\omega}$ is constant across clusters, for $\widehat{\boldsymbol{\beta}}_{OLS}$ one requires that both $\widetilde{M}_{XX,\omega}$ and $\widetilde{W}_{XX,\omega}$ are constant. Intuitively, this is because in $\widehat{\boldsymbol{\beta}}_{FE}$ the observations are transformed in terms of deviations from individual-specific averages and therefore the properties of $\widetilde{W}_{XX,\omega}$ do not matter.

[7] See Appendix A.

4 The Partitional Clustering Problem

4.1 Theory

Denote the true number of clusters by Ω_0, where $\Omega_0 < \xi$, a known constant. Thus there is a partition of the N individuals into Ω_0 clusters, $\Pi_{\Omega_0} = \{C_{0,1}, \ldots, C_{0,\Omega_0}\}$ with $C_{0,\omega}$ being the set of indices of elements in the ω^{th} cluster, $C_{0,\omega} = \{\omega_1, \ldots, \omega_{N_{0\omega}}\} \subseteq \{1, 2, \ldots, N\}$. The number of individuals in the ω^{th} cluster is $|C_{0,\omega}| = N_{0\omega}$, $N_{01} + \ldots + N_{0\Omega_0} = N$.

Each cluster has its own regression structure so

$$\mathbf{Y}_{C_{0,\omega}} = X_{C_{0,\omega}}\boldsymbol{\beta}_{0\omega} + \kappa_{C_{0,\omega}} + \boldsymbol{\eta}_{C_{0,\omega}} + \varepsilon_{C_{0,\omega}}, \text{ for } \omega = 1, ..., \Omega_0, \tag{13}$$

where $\mathbf{Y}_{C_{0\omega}} = \left(\mathbf{y}'_{\omega_1}, ..., \mathbf{y}'_{\omega_{N_{0\omega}}}\right)'$, with $\mathbf{y}_{\omega_i} = (y_{\omega i1}, ..., y_{\omega iT})'$, is the $(N_{0\omega}T) \times 1$ vector of observations on the dependent variable for the individuals in the ω^{th} cluster and $\boldsymbol{\eta}_{C_{0,\omega}} = \left(\boldsymbol{\iota}'_T\eta_{\omega_1}, ..., \boldsymbol{\iota}'_T\eta_{\omega_{N_{0\omega}}}\right)'$. $\boldsymbol{\beta}_{0\omega}$ is a vector of fixed coefficients specific to each cluster. $X_{C_{0,\omega}}$is the corresponding $(N_{0\omega}T) \times K$ matrix of covariates.

Premultiplying (13) by $Q_{C_{0,\omega}} = I_{N_{0\omega}T} - I_{N_{0\omega}} \otimes \frac{1}{T}\boldsymbol{\iota}_T\boldsymbol{\iota}'_T$ to remove the time-invariant effects yields

$$Q_{C_{0,\omega}}\mathbf{Y}_{C_{0,\omega}} = Q_{C_{0,\omega}}X_{C_{0,\omega}}\boldsymbol{\beta}_{0\omega} + Q_{C_{0,\omega}}\boldsymbol{\varepsilon}_{C_{0,\omega}}, \tag{14}$$

where I_j is a $j \times j$ identity matrix, or

$$\widetilde{\mathbf{Y}}_{C_{0,\omega}} = \widetilde{X}_{C_{0,\omega}}\boldsymbol{\beta}_{0\omega} + \widetilde{\boldsymbol{\varepsilon}}_{C_{0,\omega}}, \tag{15}$$

where $\widetilde{\mathbf{Y}}_{C_{0,\omega}} = Q_{C_{0,\omega}}\mathbf{Y}_{C_{0,\omega}}$, and so on.

The following results are motivated by the approach taken in Shao and Wu (2005) for the cross-sectional regression case. The argument is more complex in the panel data case due to the lack of independence of the terms in $\widetilde{\boldsymbol{\varepsilon}}_{C_{0\omega}}$. We make the following assumptions needed to establish the clustering result.

First, note that BA.3 ensures that for the true partition there exist fixed constants $d_\omega \in (0, 1)$ such that $d_\omega < \frac{N_{0\omega}}{N} < 1$, $\omega = 1, ..., \Omega_0$ for N large enough. We strengthen BA.1 to

CA.1 Given the covariates $X_{\omega it}$ corresponding to the observations in the ω^{th} cluster, the error vectors $\boldsymbol{\varepsilon}_{\omega i} = (\varepsilon_{\omega i1}, ..., \varepsilon_{\omega iT})'$ for the individuals in the cluster are independent and identically distributed random vectors with mean vector $\mathbf{0}$ and for some $\delta > 0$, $E\left|\varepsilon_{\omega it}\right|^{2+\delta} < \infty$. To avoid trivialities assume some elements of $\boldsymbol{\varepsilon}_{\omega i}$ have non-zero variance.

Let C_ℓ denote a true class or a subset of a true class with N_ℓ elements. Given the matrix $\widetilde{X}_{C_\ell}$, let $\widetilde{X}_{C_\ell}^{(t)}$ be the submatrix consisting of rows $t, t+T, ..., t+(N_\ell - 1)T$ of $\widetilde{X}_{C_\ell}$ for $t = 1, ..., T$.

CA.2 There exist constants $\alpha_1 > 0$ and $\alpha_2 > 0$ such that the eigenvalues of $N_\ell^{-1}\widetilde{X}_{C_\ell}'\widetilde{X}_{C_\ell}$ and $N_\ell^{-1}\widetilde{X}_{C_\ell}^{(t)\prime}\widetilde{X}_{C_\ell}^{(t)}$ lie in $[\alpha_1, \alpha_2]$ for N_ℓ large enough.

CA.3 For any column vector $\mathbf{x}_{\omega\ell}$ of $\widetilde{X}_{C_\ell}$, its elements $x_{\omega\ell}^{(1)}, ..., x_{\omega\ell}^{(N_\ell T)}$ satisfy the condition

$$\sum_{i=1}^{N_\ell T} \left| x_{\omega\ell}^{(i)} \right|^{2+\delta} = O\left[\left(\mathbf{x}_{\omega\ell}'\mathbf{x}_{\omega\ell}\right)^{(2+\delta)/2} / \log\left(\mathbf{x}_{\omega\ell}'\mathbf{x}_{\omega\ell}\right)^{1+\delta} \right] \tag{16}$$

for $1 \leq \omega \leq \Omega$ and some $\delta > 0$.

Assumptions CA.2$-$CA.3 describe the behaviour of the covariates and they will hold if, for example, the covariate vectors $\mathbf{x}_1, ..., \mathbf{x}_N$ are *i.i.d.* with appropriate moment conditions. For any set C_ℓ which is a true class, a subset of a true class or a union of subsets of a true class with $|C_\ell| = N_\ell$, let $P_{\widetilde{X}_{C_\ell}}$ denote the projection matrix

$$P_{\widetilde{X}_{C_\ell}} = \widetilde{X}_{C_\ell}\left(\widetilde{X}_{C_\ell}'\widetilde{X}_{C_\ell}\right)^{-1}\widetilde{X}_{C_\ell}', \tag{17}$$

based on the corresponding $\widetilde{X}_{C_\ell}$ matrix. Let $\boldsymbol{\varepsilon}_{C_\ell}$ denote the vector of corresponding error terms. The following lemma controls the rate of growth of a weighted sum of random variables.

Lemma 1 *Let $\varpi_1, \varpi_2, ...$ be a sequence of random variables with zero mean, such that $0 < E\left(\varpi_i^2\right) = \sigma_i^2$ and $E\left|\varpi_i\right|^{2+\delta} < \tau < \infty$ for some $\tau > 0$, $\delta > 0$ and $i = 1, 2, ...$ Furthermore, let $\alpha_1, \alpha_2, ..., \in \mathbb{R}$ be a sequence of constants such that*

$$\begin{aligned}(i) \quad B_N^2 \quad &= \quad \sum_{i=1}^N \alpha_i^2 \to \infty;\\ (ii) \quad \sum_{i=1}^N |\alpha_i|^{2+\delta} \quad &= \quad O\left\{B_N^{2+\delta}\left(\log B_N^2\right)^{-1-\delta}\right\}, \textit{ for some } \delta > 0.\end{aligned}$$

Then, almost surely, for $N \to \infty$

$$T_N = \sum_{i=1}^N \alpha_i \varpi_i = O\left(\left(B_N^2 \log\log\left(B_N^2\right)\right)^{\frac{1}{2}}\right).$$

Proof. See Shao and Wu (2005), Lemma 3.5. ■

Write

$$\boldsymbol{\varepsilon}_{C_\ell} = \boldsymbol{\nu}_{C_\ell}^{(1)} + ... + \boldsymbol{\nu}_{C_\ell}^{(T)}, \tag{18}$$

where the i^{th} element of $\boldsymbol{\nu}_{C_\ell}^{(t)}$ is $\left(\boldsymbol{\varepsilon}_{C_\ell}\right)_i I\left(i \in \{t, t+T, T+2T, ...\}\right)$. For example, $\boldsymbol{\nu}_{C_\ell}^{(1)} = \left(\varepsilon_{\omega 11}, 0, ..., 0, \varepsilon_{\omega 21}, 0, ..., \varepsilon_{\omega N_\ell 1}, 0, ..., 0\right)'$ and so on. The non-zero elements of the vector $\boldsymbol{\nu}_{C_\ell}^{(t)}$ are the *i.i.d.* error terms corresponding to the observations at time t for the elements in the cluster. We can write

$$\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = \sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}. \tag{19}$$

Using the idempotent nature of the matrix $P_{\widetilde{X}_{C_\ell}}$ and the Cauchy-Schwartz inequality we have

$$\begin{aligned} \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)^2 &= \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P^2_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)^2 \\ &= \left(\left(P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)}\right)' \left(P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)\right)^2 \\ &\leq \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)}\right) \left(\boldsymbol{\nu}_{C_\ell}^{(s)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right). \end{aligned} \tag{20}$$

Thus, if $\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)} = O\left(\log\log N_{C_\ell}\right)$ a.s. for each t, then $\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = O\left(\log\log N_{C_\ell}\right)$ a.s.

Applying Lemma 1 along with assumptions CA.1-CA.3 we have

$$\nu_{C_\ell}^{(t)\prime} \widetilde{X}_{C_\ell} = O\left(N_\ell \log\log N_\ell\right)^{\frac{1}{2}} \text{ a.s.} \tag{21}$$

Therefore,

$$\boldsymbol{\varepsilon}'_{C_\ell} \widetilde{X}_{C_\ell} = O\left(\left(N_\ell \log\log N_\ell\right)^{\frac{1}{2}}\right) \text{ a.s.} \tag{22}$$

Furthermore, CA.2 ensures that the elements of $\left(\widetilde{X}'_{C_\ell}\widetilde{X}'_{C_\ell}\right)^{-1}$ are $O\left(N_\ell^{-1}\right)$. Hence, using (21) and arguing as in the proof of Lemma A.2 of Bai, Rao and Wu (1999) we have

$$\begin{aligned} \boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)} &= \boldsymbol{\nu}_{C_\ell}^{(t)\prime} \widetilde{X}_{C_\ell} \left(\widetilde{X}'_{C_\ell}\widetilde{X}_{C_\ell}\right)^{-1} \widetilde{X}'_{C_\ell} \boldsymbol{\nu}_{C_\ell}^{(t)} \\ &= O\left(\log\log N_\ell\right). \end{aligned} \tag{23}$$

As a result,

$$\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{X}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = O\left(\log\log N_\ell\right). \tag{24}$$

The results in (22) and (24) are key to proving that the clustering algorithm converges to the true number of clusters. Using the class-growing sequence approach, the sequence of true classifications of $\{1, 2, ..., N\}$ is naturally nested as N increases, i.e.

$$C_{0,\omega}^{(N)} \subseteq C_{0,\omega}^{(N+1)} \text{ for all } \omega = 1, ..., \Omega_0, \text{ for large } N. \tag{25}$$

Consider all class-growing sequences of classifications with Ω clusters, $\Pi_{\Omega}^{(N)} = \left\{C_{\Omega 1}^{(N)}, ..., C_{\Omega\Omega}^{(N)}\right\}$, then $C_{\Omega\omega}^{(N)} \subseteq C_{\Omega\omega}^{(N+1)}$, $\omega = 1, ..., \Omega$, for large N. Let $\widehat{\boldsymbol{\beta}}_{\Omega\omega}$ be the least squares estimate of $\boldsymbol{\beta}$ based on the observations in the cluster $C_{\Omega\omega}$, $\widehat{\boldsymbol{\beta}}_{0\omega}$ be the least squares estimate of $\boldsymbol{\beta}$ based on the observations in the true cluster $C_{0,\omega}$ and $\widehat{\boldsymbol{\beta}}_{\omega|j}$ be the least squares estimate based on the observations in the cluster $C_{\Omega\omega} \cap C_{0,j}$, $\omega = 1, ..., \Omega$, $j = 1, ..., \Omega_0$.

Let

$$RSS_\omega = \left\| \widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}} \widehat{\boldsymbol{\beta}}_{\Omega\omega} \right\|^2$$

denote the sum of the squares of the fixed effect residuals for the C_{Ω_ω} cluster and let

$$RSS_T = RSS_T(\Omega) = \sum_{\omega=1}^{\Omega} RSS_\omega.$$

We use the following model information-based criterion (MIC) as the basis for determining the underlying cluster structure:

$$F_N\left(\Pi_\omega^{(N)}\right) = N \log\left(\frac{RSS_T}{NT}\right) + f(\Omega)\,\theta_N, \tag{26}$$

where $f(\Omega)$ is a strictly increasing function of Ω and θ_N is a sequence of constants. Using this criterion to compare two distinct partitions we have, recalling
$\log(1+x) \sim x$ for small x,

$$\begin{aligned} & F_N\left(\Pi_\Omega^{(N)}\right) - F_N\left(\Pi_{\Omega_0}^{(N)}\right) \\ = \; & N\log\left(\frac{RSS_T(\Omega)}{NT}\right) - N\log\left(\frac{RSS_T(\Omega_0)}{NT}\right) \\ & +[f(\Omega) - f(\Omega_0)]\,\theta_N, \\ = \; & N\log\left[1 + \frac{RSS_T(\Omega) - RSS_T(\Omega_0)}{RSS_T(\Omega_0)}\right] \\ & +[f(\Omega) - f(\Omega_0)]\,\theta_N, \\ \sim \; & \frac{RSS_T(\Omega) - RSS_T(\Omega_0)}{RSS_T(\Omega_0)/N} + [f(\Omega) - f(\Omega_0)]\,\theta_N. \end{aligned}$$

The residual sum of squares for the Π_{Ω_0} partition divided by N is a measure of the variability in the data. Thus, heuristically, the first term compares the goodness of fit of the models normed by a measure of the overall level of spread, while the second term is a penalty for overfitting.

Let $\widehat{\Omega}_0$ be the estimate of Ω_0 that minimises F_N, i.e.

$$F_N\left(\Pi_{\widehat{\Omega}_0}^{(N)}\right) = \min_{1 \le \Omega \le \xi} \min_{\Pi_\Omega^{(N)}} F_N\left(\Pi_\Omega^{(N)}\right). \tag{27}$$

The following theorem shows that the criterion in (26) selects the true number of clusters amongst all class-growing sequences with probability one for N large enough.

Theorem 2 *Let* $\lim_{N\to\infty} N^{-1}\theta_N = 0$ *and* $\lim_{N\to\infty} (\log\log N)^{-1}\theta_N = \infty$*. Suppose that assumptions CA.1-CA.3 and BA.4 hold and* Π_{Ω_0} *is the true clustering partition corresponding to model* (13)*. Then the MIC criterion is strongly consistent – that is, it selects* Ω_0*, the true number of clusters among all class-growing sequences, with probability one as* $N \to \infty$*.*

Proof. See Appendix B. ■

The first condition in the above theorem prevents estimating too many clusters asymptotically while the second conditions prevents under-fitting. Similar conditions underlie well-known model selection criteria such as the AIC and the BIC that are often used in standard model selection. The notable difference is that our theorem is developed for the purpose of clustering individuals and therefore the asymptotics are implemented via class-growing sequences.

The above result in fact carries across to the more general model

$$y_{\omega it} = \boldsymbol{\beta}'_{\omega}\mathbf{x}_{\omega it} + u_{\omega it}, \;\; u_{\omega it} = \boldsymbol{\lambda}'_{\omega i}\boldsymbol{\phi}_t + \varepsilon_{\omega it}, \tag{28}$$

where $\boldsymbol{\lambda}_{\omega i} = (\lambda_{1\omega i}, ..., \lambda_{P\omega i})'$, $\boldsymbol{\phi}_t = (\phi_{1t}, ..., \phi_{Pt})'$ are both $P \times 1$ vectors and give rise to a multi-factor error structure. This can be useful to characterise individual-specific unobserved heterogeneity, captured by $\boldsymbol{\lambda}_{\omega i}$, which varies over time, and also allow for the existence of common unobserved shocks (such as technological shocks and financial crises), captured by $\boldsymbol{\phi}_t$, the impact of which is different for each individual i. The model in (28) reduces back to the structure in (13) by setting $P = 1$ and $\phi_t = 1$ for all t. Writing (28) in vector form we have

$$\mathbf{Y}_{C_{0,\omega}} = X_{C_{0,\omega}}\boldsymbol{\beta}_{0\omega} + (I_{N_{0\omega}} \otimes \Phi)\,\boldsymbol{\lambda}_{C_{0,\omega}} + \varepsilon_{C_{0,\omega}}, \tag{29}$$

where $\Phi = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_T)'$ is a $T \times P$ matrix and $\boldsymbol{\lambda}_{C_{0,\omega}} = \left(\boldsymbol{\lambda}'_{\omega 1}, ..., \boldsymbol{\lambda}'_{\omega N_{0\omega}}\right)'$ is a $N_{0\omega}P \times 1$ vector. By pre-multiplying the model above by the matrix $M_{C_{0,\omega}} = I_{N_{0\omega}T} - I_{N_{0\omega}} \otimes \left[\Phi\left(\Phi'\Phi\right)^{-1}\Phi'\right]$ the model reduces to a classical form similar to (15). Thus, interpreting $\widetilde{X}_{C_l}$ in the various conditions as $M_{C_l}X$, we have the following corollary:

Corollary 3 *Let* $\lim_{N\to\infty} N^{-1}\theta_N = 0$ *and* $\lim_{N\to\infty} (\log\log N)^{-1}\theta_N = \infty$*. Suppose that assumptions CA.1-CA.3 and BA.4 hold and* Π_{Ω_0} *is the true clustering* *partition corresponding to model* (28)*. Then the MIC criterion is strongly consistent – that is, it selects Ω_0, the true number of clusters among all class-growing sequences, with probability one as $N \to \infty$.*

In practice Φ is unknown and needs to be replaced with a consistent estimate. This can be obtained using for example principal components analysis (see e.g. Connor and Korajzcyk, 1986, and Bai, 2003) or the method of Pesaran (2006).

4.2 Implementation

The number of ways to partition a set of N objects into Ω nonempty subsets is given by a 'Stirling number of the second kind', which is one of two types of Stirling numbers that commonly occur in the field of combinatorics.[8] Stirling numbers of the second kind are given by the formula

$$\begin{aligned} S(N,\Omega) &= \sum_{\omega=1}^{\Omega} (-1)^{\Omega-\omega} \frac{\omega^{N-1}}{(\omega-1)!\,(\Omega-\omega)!} \\ &= \frac{1}{\Omega!} \sum_{\omega=0}^{\Omega} (-1)^{\Omega-\omega} \binom{\Omega}{\omega} \omega^{N}. \end{aligned} \tag{30}$$

The total number of ways to partition a set of N objects into non-overlapping sets is given by the N^{th} Bell number

$$B_N = \sum_{\Omega=1}^{N} S(N,\Omega). \tag{31}$$

To see the order of the magnitude of a Stirling number, for $N = 50$ and $\Omega = 2$ the total number of distinct partitions is larger than $5.6 \cdot 10^{14}$. This implies that if we assumed, rather optimistically, that a given computer was able to estimate $10,000$ panel regressions every second, then one would require about 1790 years to exhaust all possible partitions. Clearly, a global search over all possible partitions is not feasible, even with small data sets – a problem that also applies to procedures based on hierarchical clustering, of course. To deal with this issue, we apply a hill-climbing algorithm of the kind used in standard partitional cluster analysis (see, e.g., Everitt, 1993). The algorithm we adopt in this paper can be outlined in the following steps[9]:

[8] See, for example, Rota (1964).

[9] The algorithm is written as an ado file in Stata 11 and it will available to all Stata users on the web.

1. Given an initial partition and a fixed number of clusters, run the fixed effects estimator for each cluster separately and calculate RSS_T;

2. Re-allocate the i^{th} cross-section to all remaining clusters and obtain the resulting RSS_T that arises each time. Finally, allocate the i^{th} individual into the cluster that achieves the smaller RSS_T;

3. Repeat the same procedure for $i = 1, ..., N$;

4. Repeat steps 2-3 until RSS_T cannot be minimised any further.

5. Once the partition that achieves the minimum value of RSS_T has been determined, repeat steps 1-4 for different number of clusters;

6. Pick the number of clusters that minimises
$$N \log \left(\frac{RSS_T}{NT} \right) + f\left(\Omega\right) \theta_N, \tag{32}$$
where $f\left(\Omega\right)$ is a strictly increasing function of Ω and θ_N is chosen such that it satisfies the bounds in Theorem 2.

Since only a local search over different partitions is feasible, the final outcome of such algorithms might be sensitive to the way of choosing the initial partition. One way to handle this issue is to select the initial partition on the basis of the slope coefficients estimated for each individual using the T observations. A different solution would be to select the initial partition based on observed attributes. Alternatively, one may use several random starts to identify locally optimal solutions and pick up the one that corresponds to the minimum RSS_T.[10] Once the number of clusters, together with the corresponding partition, has been determined, it is also possible to cross-validate the results. Several methods have been proposed in the clustering literature for this purpose. One simple approach involves perturbating the observations and checking whether the clusters are robust with respect to changes in the data. Kaufman and Rousseeuw (2005) describe alternative methods.

5 Simulation Study

In this section we carry out a simulation experiment to investigate the performance of our criterion in finite samples. Our main focus is on the choice of θ_N and the

[10] This is the practice adopted in our simulation experiment that follows.

effect of (i) the number of clusters, (ii) the size of N, (iii) the number of regressors and (iv) the signal-to-noise ratio in the model. We also pay attention to the properties of the estimators that arise from the estimated partitions, as well as the pooled fixed effects and OLS estimators.

5.1 Experimental Design

The underlying process is given by

$$\begin{aligned} y_{\omega it} &= \sum_{k=1}^{K} \beta_{k\omega} x_{k\omega it} + \eta_{\omega i} + \varepsilon_{\omega it}, \\ t &= 1,...,T,\ i\,[\in \omega] = 1,...,N_{\omega} \text{ and } \omega = 1,...,\Omega_0, \end{aligned} \tag{33}$$

where $\eta_{\omega i}$ and $\varepsilon_{\omega it}$ are drawn in each replication from $i.i.d.N\left(0,\sigma_{\eta}^2\right)$ and $i.i.d.N\left(0,\sigma_{\varepsilon_\omega}^2\right)$ respectively, while $x_{k\omega it}$ is drawn from $i.i.d.N\left(\mu_{x_{k\omega}},\sigma_{x_{k\omega}}^2\right)$.

Define $y_{\omega it}^{*} = y_{\omega it} - \eta_{\omega i}$, such that (33) can be rewritten as

$$y_{\omega it}^{*} = \sum_{k=1}^{K} \beta_{k\omega} x_{k\omega it} + \varepsilon_{\omega it}, \tag{34}$$

and let the signal-to-noise ratio be denoted by $\zeta_\omega = \sigma_{s_\omega}^2/\sigma_{\varepsilon_\omega}^2$, where $\sigma_{s_\omega}^2$ and $\sigma_{\varepsilon_\omega}^2$ denote the variance of the signal and noise, respectively, for the ω^{th} cluster. $\sigma_{s_\omega}^2$ equals

$$\sigma_{s_\omega}^2 = var\left(y_{\omega it}^{*} - \varepsilon_{\omega it}\right) = var\left(\sum_{k=1}^{K} \beta_{k\omega} x_{k\omega it}\right) = \sum_{k=1}^{K} \beta_{k\omega}^2 \sigma_{x_{k\omega}}^2. \tag{35}$$

This implies that for a given value of $\left\{\sigma_{x_{k\omega}}^2\right\}_{k=1}^{K}$ and $\sigma_{\varepsilon_\omega}^2$, the signal-to-noise ratio for the ω^{th} cluster depends on the value of $\{\beta_{k\omega}\}_{k=1}^{K}$. Thus, for example, scaling the coefficients upwards by a constant factor will increase ζ and this may improve the performance of the model selection criterion; however, there is no natural way to choose the value of such scalar. Furthermore, notice that for fixed $\sigma_{\varepsilon_\omega}^2$ alternating K will change $\sigma_{s_\omega}^2$ and thereby the performance of the criterion may also be affected. We control both these effects by normalising $\sigma_{\varepsilon_\omega}^2 = 1$, $\zeta_\omega = \zeta$, for $\omega = 1,...,\Omega_0$ and setting $\sigma_{x_{k\omega}}^2 = \zeta/\left(\beta_{\omega k}^2 K\right)$. In this way, the signal-to-noise ratio in our design is invariant to the choice of K and the scale of $\{\beta_{k\omega}\}_{k=1}^{K}$. The values of the slope coefficients are listed in Table 1. We consider $\zeta = \{4, 8\}$, $N = \{100, 400\}$ with $T = 10$, $K = \{1, 4\}$ and $\Omega_0 = \{1, 2, 3\}$.[11] We set $N_1 = 0.7N$, $N_2 = 0.3N$ for $\Omega_0 = 2$ and $N_1 = 0.4N$, $N_2 = 0.3N$, $N_2 = 0.3N$ for $\Omega_0 = 3$. This allows the size of the clusters to be different. We perform 500 replications in each

[11] We also set $\mu_{x_{k\omega}} = 1$ for $k = 1,...,K$ and $\omega = 1,...,\Omega$.

experiment. To reduce the computational burden, we fit models with $\Omega = 1, 2, 3$ clusters when $\Omega_0 = 1$, $\Omega = 1, 2, 3, 4$ clusters when $\Omega_0 = 2$ and $\Omega = 1, 2, 3, 4, 5$ clusters when $\Omega_0 = 3$.

To examine the performance of the criterion under a factor error structure, we set up an additional design where now

$$\varepsilon_{\omega it} = \lambda_{\omega i}\phi_t + \upsilon_{\omega it}, \tag{36}$$

where $\lambda_{\omega i} \sim i.i.dU[-1, 1]$, $\phi_t \sim i.i.d.N(0, 1)$ and $\upsilon_{\omega it} \sim i.i.d.N(0, 1)$. In this case, ϕ_t is estimated in each replication from the vector of principal components extracted from $y_{\omega it}$ and the model is orthogonalised prior to estimation by premultiplying the $T \times 1$ vectors of observed variables, $\mathbf{y}_{\omega i} = (y_{\omega i1}, ..., y_{\omega iT})'$ and $\mathbf{x}_{k\omega i} = (x_{k\omega i1}, ..., x_{k\omega iT})'$ for $k = 1, ..., K$, by the $T \times T$ idempotent matrix $M = I_T - H(H'H)^{-1}H'$, $H = \left(\widehat{\phi}_1, ..., \widehat{\phi}_T\right)'$.

Table 1. Parameter values used in the simulation study.

	$K = 1$	$K = 4$
$\Omega_0 = 1$	$\beta = 1$	$\boldsymbol{\beta} = \begin{pmatrix} 1 \\ .5 \\ .75 \\ 2 \end{pmatrix}$
$\Omega_0 = 2$	$\beta_1 = 1$ $\beta_2 = .5$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 1 \\ .5 \\ .75 \\ 2 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} .5 \\ .25 \\ .375 \\ 1 \end{pmatrix}$
$\Omega_0 = 3$	$\beta_1 = 1$ $\beta_2 = .5$ $\beta_3 = -.25$	$\boldsymbol{\beta}_1 = \begin{pmatrix} 1 \\ .5 \\ .75 \\ 2 \end{pmatrix}, \boldsymbol{\beta}_2 = \begin{pmatrix} .5 \\ .25 \\ .375 \\ 1 \end{pmatrix}, \boldsymbol{\beta}_3 = \begin{pmatrix} -.25 \\ 1 \\ 1.5 \\ 0.5 \end{pmatrix}$

5.2 Results

Tables A1-A3 in the appendix report the results of our simulation experiments in terms of the relative frequency of selecting Ω clusters when the true number of clusters is Ω_0. The relative frequency of selecting the true number of clusters is emphasised in bold. Since the property of consistency of $\widehat{\Omega}$ only requires that $f(\Omega)$ is stricly increasing in Ω and θ_N satisfies $\lim_{N\to\infty} N^{-1}\theta_N = 0$ and $\lim_{N\to\infty} (\log\log N)^{-1}\theta_N = \infty$, there is a broad range of values one can choose. In this study we set $f(\Omega) = \Omega$ such that

$$MIC_j = N \log\left(\frac{RSS_T}{NT}\right) + \Omega\theta_j \text{ for } j = 1, ..., 4,$$

where $\theta_1 = 2$, $\theta_2 = \log N$, $\theta_3 = (1/\psi)\left[(\log N)^{\psi} - 1\right]$ and $\theta_4 = \sqrt{N}$. MIC_1 and MIC_2 resemble the Akaike and Bayesian information criteria, respectively, except that they are applied to the clustering selection problem The choice for θ_3 is motivated from the fact that $(1/\psi)\left[(\log N)^{\psi} - 1\right] \to \log\log N$ as $\psi \to 0$ and hence for any ψ sufficiently large, the lower bound of Theorem 2 is satisfied.[12] Notice that when $N = 100$, $\theta_3 \approx \theta_2$ for $\psi = 3$ and $\theta_3 \approx \theta_4$ for $\psi = 6$. Therefore, we set $\psi = 4.5$.

As we can see from the tabulated results, MIC_1 performs poorly in most circumstances in that it constantly overestimates the true number of clusters. This is not surprising as the criterion is not consistent. In fact, its performance deteriorates as N increases. MIC_2 is a special case of our criterion and performs somewhat better than MIC_1. Notwithstanding, in a lot of cases it largely overestimates the number of clusters, especially when $\Omega_0 = 1, 2$. We have explored further the underlying reason for this result. We found that a larger penalty is required in the clustering regression problem to prevent over-fitting than what is typically used in the standard model selection problem. On the other hand, both MIC_3 and MIC_4 perform very well in all circumstances. This holds true for all values of N, K and ζ. Naturally, the performance of both criteria improves with larger values of N and ζ.[13]

Similar conclusions can be drawn from the model with the factor error structure, the results of which are reported in Tables A1(b)-A3(b) in the Appendix. In particular, while MIC_1 and MIC_2 tend to overestimate the true number of clusters, as before, MIC_3 continues to perform very well even for $\zeta = 4$ and $N = 100$. The performance of MIC_4 slightly deteriorates in this case, although it behaves differently to MIC_1 and MIC_2. In particular, since MIC_4 has the largest penalty for overfitting it errs on the side of underestimating the true number of clusters. All in all, the simulation results show that MIC_3 and MIC_4, especially the former, perform very well under a variety of parametrizations.

Tables A4 and A4(b) in the appendix report the average point estimates of the parameters for $K = 1$.[14] Standard deviations are reported in parentheses. 'Pooled FE' and 'pooled OLS' denote the fixed effects and OLS estimates that arise by pooling all clusters together, i.e. ignoring cluster-specific heterogeneity in

[12] See e.g. Shao and Wu (2005).

[13] ζ does not affect the results when $\Omega_0 = 1$ of course.

[14] To save space, we do not report the results obtained for $K = 4$ because similar conclusions can be drawn.

the slope parameters. FE_ω denotes the fixed effects estimate of the parameter for the ω^{th} cluster that arises from the estimated partition when Ω_0 is estimated using MIC_3. For the pooled estimators, the true coefficient is taken to be the weighted average value of the cluster-specific unknown slope coefficients, with the weights determined by the size of the clusters. As we can see, the bias of both pooled FE and pooled OLS is large and becomes more alike as N increases. The negative direction of the bias is due to the fact that the clusters with smaller coefficients exhibit relatively larger leverage because the variance of the regressors is larger for these clusters. On the other hand, the cluster-specific fixed effects estimators are virtually unbiased even if they are obtained from estimated clusters and the corresponding estimated partitions. This holds true even for $N = 100$, although the performance of the estimators naturally improves as N increases. In conclusion, we see that the criterion performs well, not only with respect to the estimate of Ω_0, but also in terms of leading to accurate cluster-specific coefficients.

6 Empirical Example

As an illustration of the proposed clustering method, we estimate a partially heterogeneous cost function using a panel data set of commercial banks operating in the United States. The issue of how to estimate scale economies and efficiency in the banking industry has attracted considerable attention among researchers due to the significant role that financial institutions play in economic prosperity and growth and, as a result, the major implications that these estimates entail for policy making.

6.1 Existing Evidence

In an earlier survey conducted by Berger and Humphrey (1997), the authors report more than 130 studies focusing on the measurement of economies of scale and the efficiency of financial institutions in 21 countries. They conclude that while there is lack of agreement among researchers regarding the preferred model with which to estimate efficiency and returns to scale, there seems to be a consensus on the fact that the underlying technology is likely to differ among banks. To this end, McAllister and McManus (1993) argue that the estimates of the returns to scale in the banking industry may be largely biased if one applies a single cost function to the whole sample of banks. This result is likely to remain even if one uses a more flexible functional form in the data, such as the translog form, because this would restrict, for example, banks of different size to share the same symmetric average cost curve. Hence, other interesting possibilities would be precluded, such as flat segments in the average cost curve over some ranges, or even different average cost curves among banks, depending on their size. Thus, the authors conclude:

> "These results, taken together, suggest that estimated cost functions vary substantially depending on the range of bank sizes included in the sample. This extreme dependence of the results on the choice of the sample suggests that there are difficulties with the statistical techniques employed", page 389.

Similarly, Kumbhakar and Tsionas (2008) argue that since the banking industry contains banks of vastly different size, the underlying technology is very likely to be different across banks:

> "The distribution of assets across banks is highly skewed. As a result of this, it is very likely that the parameters of the underlying technology (cost function in this case) will differ among banks", page 591.

Since this view appears to have been widely adopted in the banking literature, we estimate a partially heterogeneous cost regression model. A similar approach conceptually has been followed indirectly by Kaparakis et al (1994), who distinguish between small and large banks and partition the population into two equally-sized sub-samples. However, this partioning is rather arbitrary and there is no formal justification for imposing two clusters.

6.2 The Data Set

The data set consists of a random sample of 551 banks, each observed over a period of 15 years. These data have been collected from the electronic database maintained by the Federal Deposit Insurance Corporation (FDIC).[15] The relatively large size of N implies that the practice of restricting the slope coefficients to be homogeneous across the whole sample may not be warranted, while the small size of T prohibits estimating a separate cost function for each individual bank in a meaningful way.

[15] See http://www.fdic.gov

6.3 Specification of Cost, Outputs and Input Prices

In the theory of banking there is not a univocal approach regarding one's view of what banks produce and what purposes they serve. In this paper we follow the "intermediation" approach, in which the banks are viewed as intermediators of financial and physical resources and produce loans and investments. Under this approach, outputs are measured in money values and cost figures include interest expenses. The selection of inputs and outputs follows closely the study conducted by Hancock (1986). The variables used in the analysis are: c; the sum of the cost related to the three input prices that appear below, y_1; the sum of industrial, commercial and individual loans, real estate loans and other loans and leases, y_2; all other assets, p_l; the price of labour, measured as total expenses on salaries and employee benefits, divided by the total number of employees, p_k; the price of capital, measured as expenses on premises and equipment, divided by the dollar value of premises and equipment, and p_f; the price of loanable funds, measured as total expenses on interest, divided by the dollar value of deposits, federal funds purchased and other borrowed funds.

Hence, the model is specified as follows[16]:

$$\begin{aligned} c_{\omega it} &= \beta_{1\omega} y_{1,\omega it} + \beta_{2\omega} y_{2,\omega it} + \gamma_{1\omega} p_{l,\omega it} + \gamma_{2\omega} p_{k,\omega it} + \gamma_{3\omega} p_{f,\omega it} + u_{\omega it}, \\ u_{\omega it} &= \alpha_{\omega i} + \varepsilon_{\omega it}. \end{aligned} \tag{37}$$

6.4 Results

We cluster the sample of banks into up to six clusters based on the algorithm developed in Section 4. The initial partition is chosen on the basis of bank size, which is proxied by the fifteen-year average value of total assets for each individual bank. Table 2 reports the values of MIC_j, $j = 1, ..., 4$, for $\Omega = 1, ..., 6$. As we can see, both MIC_3 and MIC_4 suggest the presence of four clusters. On the other hand, the performance of MIC_2 appears to be similar to MIC_1 in that they both return high scores for Ω_0. This is consistent with the results of the simulation study, which show that a larger penalty is required in the clustering regression problem to prevent over-fitting compared to the standard model selection problem.

[16] All variables are in logs.

Table 2. Results for estimating the number of clusters.

Ω	1	2	3	4	5	6
MIC_1	-717.9	-918.5	-958.6	-985.4	-995.8	**-1014.4**
MIC_2	-717.1	-917.0	-956.3	-982.4	-992.1	**-1010.0**
MIC_3	-699.3	-881.4	-902.9	**-911.2**	-903.2	-903.1
MIC_4	-696.4	-875.6	-894.1	**-899.1**	-888.5	-885.6

Table 3 reports the estimation results obtained for model (37) when $\Omega = 4$. We adopt a notation similar to the simulation study; in particular, pooled FE (OLS) denotes the FE (OLS) estimate for the sample as a whole, FE_j refers to the fixed effects estimate for the j^{th} cluster and $\overline{FE}$ is the weighted average FE estimate of all clusters with the weights determined by the size of each cluster. The clusters are sorted in ascending order such that cluster 1 contains on average the smallest banks and cluster 4 the largest banks.

We can see that there are some large and statistically significant differences in the value of the coefficients across clusters. For example, the estimated coefficient of the price of labour, $\widehat{\gamma}_1$, appears to be strictly decreasing in the size of the banks. This might be explained by the fact that large banks usually make larger in value loans while the labour cost for an individual loan is the same, which implies that the labour cost component tends to be smaller for large loans. On the other hand, the estimated coefficient of loans, $\widehat{\beta}_1$, appears to rise as bank size increases, although it remains well below one. This implies that while there are increasing output returns for both small and large banks, the benefit of small banks getting larger is higher than for banks which are already large. In general, we see that banks of different size have different cost drivers and therefore pooling the data and imposing homogeneity in the slope parameters across the whole sample may yield misleading results. This becomes apparent when we compare pooled FE with $\overline{FE}$, the difference of which is statistically significant for most coefficients.

Table 3. Estimation Results^{a,b}

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$
pooled FE	.138	.414	.270	.023	.372
	(.004)	(.005)	(.004)	(.006)	(.008)
pooled OLS	.247	.432	.257	.003*	.342
	(.003)	(.005)	(.006)	(.004)	(.010)
FE_1	.037	.036	.720	.005	.442
	(.004)	(.006)	(.005)	(.006)	(.008)
FE_2	.051	.330	.533	-.035*	.400
	(.005)	(.007)	(.007)	(.051)	(.007)
FE_3	.217	.542	.161	.002*	.377
	(.005)	(.008)	(.005)	(.009)	(.012)
FE_4	.293	.325	.121	.074	.588
	(.006)	(.007)	(.004)	(.011)	(.011)
$\overline{FE}$	.141	.298	.402	.010	.451

(a) Standard Errors in Parentheses. (b) * denotes an insignificant regressor at the 5% level.

7 Concluding Remarks

Full homogeneity versus full parameter heterogeneity is a topic that has intrigued research in the analysis of panel data over the last few decades at least. In many cases the issue remains practically unresolved; for example, Burnside (1996) rejected the hypothesis that production function parameters are homogeneous across a panel of US manufacturing industries. Similarly, Baltagi and Griffin (1997) rejected the hypothesis that gasoline demand elasticities were equal across a panel of OECD countries. Despite this, both studies found that fully heterogeneous estimators led to very imprecise estimates, which, in some cases, had even the wrong sign. This paper has proposed an intermediate modelling framework that imposes only partially heterogeneous restrictions in the parameters, based on the concept of 'partitional clustering'. The unknown number of clusters, together with the corresponding partition, is estimated using an information-based criterion that is strongly consistent for fixed T. The partitional clustering algorithm we have developed for Stata 11 is available on the web.

References

[1] Antweiler, Werner, "Nested Random Effects Estimation in Unbalanced Panel Data," *Journal of Econometrics* 101:2 (2001), 295-313.

[2] Bai, Jushan, "Inferential Theory for Factor Models of Large Dimensions," *Econometrica* 71 (2003), 135-173.

[3] Bai, Zhidong, Calyampudi R. Rao, and Yuehua Wu, "Model Selection with Data-Oriented Penalty," *Journal of Statistical Planning and Inference* 77 (1999), 103-117.

[4] Baltagi, Badi, *Econometric Analysis of Panel Data*, 4th ed. (West Sussex: John Willey & Sons, 2008).

[5] Baltagi, Badi H., and James M. Griffin, "Pooled Estimators vs. their Heterogeneous Counterparts in the Context of Dynamic Demand for Gasoline," *Journal of Econometrics* 77 (1977), 303-327.

[6] Baltagi, Badi H., James M. Griffin, and Weiwen Xiong, "To Pool or not to Pool: Homogeneous Versus Heterogeneous Estimators Applied to Cigaretter Demand," *Review of Economics and Statistics* 82:1 (2000), 117-126.

[7] Baltagi, Badi H., Georges Bresson, and Alain Pirotte, "Comparison of Forecast Performance for Homogeneous, Heterogeneous and Shrinkage Estimators. Some Empirical Evidence from US Electricity and Natural-gas Consumption," *Economics Letters* 76 (2002), 375-382.

[8] Baltagi, Badi H., Seuck H. Song, and Byoung C. Jung, "The Unbalanced Nested Error Component Regression Model," *Journal of Econometrics* 101 (2001), 357-381.

[9] Berger, Allen N., and David B. Humphrey, "Efficiency of Financial Institutions: International Survey and Directions for Future Research," *European Journal of Operational Research* 98 (1997), 175-212.

[10] Burnside, Craig, "Production Function Regressions, Returns to Scale, and Externalities," *Journal of Monetary Economics* 37 (1996), 177-201.

[11] Cameron, Colin A., and Pravin K. Trivedi, *Microeconometrics: Methods and Applications*, (New York: Cambridge University Press, 2005).

[12] Connor, Gregory, and Robert A. Korajzcyk, “Performance Measurement with the Arbitrage Pricing Theory: A New Framework for Analysis,” *Journal of Financial Economics* 15 (1986), 373-394.

[13] Durlauf, Steven, and Paul Johnson, “Multiple Regimes and Cross-country Growth Behaviour,” *Journal of Applied Econometrics* 10 (1995), 365–384.

[14] Everitt, Brian, *Cluster analysis*, 3rd ed. (London: Eward Arnold, 2003).

[15] Galor, Oded, “Convergence? Inference from Theoretical Models,” *Economic Journal* 106 (1996), 1056-1069.

[16] Hancock, Diana, “A Model of Financial Firm with Imperfect Asset and Deposit Elasticities,” *Journal of Banking and Finance* 10 (1986), 37-54.

[17] Hsiao, Cheng, *Analysis of Panel Data*, 2nd ed. (Cambridge: Cambridge University Press, 2003).

[18] Kapetanios, George, “Cluster Analysis of Panel Datasets Using Non-Standard Optimisation of Information Criteria,” *Journal of Economic Dynamics and Control* 30:8 (2006), 1389-1408.

[19] Kaparakis, Emmanuel I., Stephen M. Miller, and Athanasios G. Noulas “Short-Run Cost Inefficiency of Commercial Banks: A Flexible Frontier Approach,” *Journal of Money, Credit and Banking* 26:4 (1994), 875-893.

[20] Kaufman, Leonard, and Peter J. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, (NY: John Wiley & Sons 1990).

[21] Kumbhakar, Subal C., and Efthymios G. Tsionas, “Scale and efficiency measurement using a semiparametric stochastic frontier model: evidence from the U.S. commercial banks,” *Empirical Economics* 34 (2008), 585-602.

[22] McAllister, Patrick H., and Douglas A. McManus, “Resolving the Scale Efficiency Puzzle in Banking,” *Journal of Banking and Finance* 17 (1993), 389-405.

[23] Pesaran, Hashem M., “Estimation And Inference In Large Heterogeneous Panels With A Multifactor Error Structure,” *Econometrica* 74:4 (2006), 967-1012.

[24] Pesaran, Hashem M., Yongcheol Shin, and Ron J. Smith, "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels," *Journal of the American Statistical Association* 94 (1999), 621-634.

[25] Rota, Gian-Carlo, "The Number of Partitions of a Set," *American Mathematical Monthly* 71:5 (1964), 498-504.

[26] Shao, Qing, and Yuehua Wu, "A Consistent Procedure for Determining the Number of Clusters in Regression Clustering," *Journal of Statistical Planning and Inference* 135 (2005), 461-476.

[27] Temple, Jonathan, "The New Growth Evidence," *Journal of Economic Literature* 37:1 (1999), 112-156.

[28] Vahid, Farshid, "Partial Pooling: A Possible Answer to Pool or Not to Pool," in *Cointegration, Causality and Forecasting: Festschrift in Honor of Clive W. J. Granger*, ed. by R. Engle and H. White, 1999.

Appendices

A Proof of Equation (12)

We have

$$
\begin{aligned}
& \operatorname{plim}_{N\to\infty}\left(\widehat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}\right) = \\
= \; & \operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}X'_{\omega i}X_{\omega i}\right)\right]^{-1}\left\{\operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}X'_{\omega i}\boldsymbol{\varepsilon}_{\omega i}\right)\right]\right. \\
& \left. +\operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}X'_{\omega i}\boldsymbol{\iota}_T\eta_{\omega i}\right)\right]+\operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}X'_{\omega i}X_{\omega i}\boldsymbol{\delta}_\omega\right)\right]\right\} \\
= \; & \operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\left(X'_{\omega i}X_{\omega i}+T\overline{X}'_{\omega i}\overline{X}_{\omega i}-T\overline{X}'_{\omega i}\overline{X}_{\omega i}\right)\right)\right]^{-1} \\
& \cdot\operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\left(X'_{\omega i}X_{\omega i}+T\overline{X}'_{\omega i}\overline{X}_{\omega i}-T\overline{X}'_{\omega i}\overline{X}_{\omega i}\right)\right)\boldsymbol{\delta}_\omega\right] \\
= \; & \operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}X'_{\omega i}Q_TX_{\omega i}+\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}T\,\overline{X}'_{\omega i}\overline{X}_{\omega i}\right)\right]^{-1} \\
& \cdot\operatorname{plim}_{N\to\infty}\left[\sum_{t=1}^{T}\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}X'_{\omega i}Q_T\mathbf{X}_{\omega i}+T\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\overline{X}'_{\omega i}\overline{X}_{\omega i}\right)\boldsymbol{\delta}_\omega\right] \\
= \; & \left[\sum_{\omega=1}^{\Omega}\left(\widetilde{M}_{XX,\omega}+T\widetilde{W}_{XX,\omega}\right)c_\omega\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\left(\widetilde{M}_{XX,\omega}+T\widetilde{W}_{XX,\omega}\right)c_\omega\boldsymbol{\delta}_\omega\right]. \qquad (38)
\end{aligned}
$$

B Proof of Theorem 2

B1. Overparameterised case: $\Omega_0 < \Omega < \xi$.

Write,

$$
\begin{aligned}
& F_N\left(\Pi_{\Omega}^{(N)}\right)-F_N\left(\Pi_{\Omega_0}^{(N)}\right) \\
& \quad = \; N\log\left[1+\frac{RSS_T(\Omega)-RSS_T(\Omega_0)}{RSS_T(\Omega_0)}\right]+\left[f\left(\Omega\right)-f\left(\Omega_0\right)\right]\theta_N, \\
& \quad = \; N\left(\frac{RSS_T(\Omega)-RSS_T(\Omega_0)}{RSS_T(\Omega_0)}+o\left(\frac{RSS_T(\Omega)-RSS_T(\Omega_0)}{RSS_T(\Omega_0)}\right)\right)+\left[f\left(\Omega\right)-f\left(\Omega_0\right)\right]\theta_N.
\end{aligned}
$$

We need to show that $F_N\left(\Pi_{\Omega}^{(N)}\right)-F_N\left(\Pi_{\Omega_0}^{(N)}\right) > 0$ a.s. for large N. We know $[f\left(\Omega\right)-f\left(\Omega_0\right)] > 0$ and, under the conditions of the theorem, θ_N grows faster than $\log\log N$. Further, as $N \to \infty$, $RSS_T(\Omega_0)/N$ is bounded away from 0 and ∞ almost surely (see, for example, Lemma 2.1 Bai et al. (1999)). Thus the result follows if we can show $RSS_T(\Omega) - RSS_T(\Omega_0) = O(\log\log N)$.

Arguing as in Shao and Wu (2005), we have

$$\begin{aligned}
& RSS_T(\Omega) - RSS_T(\Omega_0) \\
= \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{X}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right\} \\
\geq \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_{0,j}} - \widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}\widehat{\boldsymbol{\beta}}_{\omega|j}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{X}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right\} \\
= \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\mathbf{Y}}'_{C_{\Omega\omega}\cap C_{0,j}}\left(I - P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0j}}}\right)\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_{0,j}} - \sum_{j=1}^{\Omega_0}\widetilde{\mathbf{Y}}'_{C_{0,j}}\left(I - P_{\widetilde{X}_{C_{0,j}}}\right)\widetilde{\mathbf{Y}}_{C_{0,j}}\right\} \\
= \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{\Omega\omega}\cap C_{0,j}}\left(I - P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}} - \sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{0,j}}\left(I - P_{\widetilde{X}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}\right\} \\
= \; & T^{-1}\left\{\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{0,j}}P_{\widetilde{X}_{C_{0,j}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} - \sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{\Omega\omega}\cap C_{0,j}}P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}\right\} \\
= \; & T^{-1}\left\{\sum_{j=1}^{\Omega_0}\boldsymbol{\varepsilon}'_{C_{0,j}}P_{\widetilde{X}_{C_{0,j}}}\boldsymbol{\varepsilon}_{C_{0,j}} - \sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\boldsymbol{\varepsilon}'_{C_{\Omega\omega}\cap C_{0,j}}P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}}\boldsymbol{\varepsilon}_{C_{\Omega\omega}\cap C_{0,j}}\right\},
\end{aligned} \tag{39}$$

where the last line follows from the idempotent nature of the matrices $Q_{C_{0,j}}$ and $Q_{C_{\Omega\omega}\cap C_{0,j}}$;

$$Q'_{C_{0,j}}P_{\widetilde{X}_{C_{0,j}}}Q_{C_{0,j}} = P_{\widetilde{X}_{C_{0,j}}}. \tag{40}$$

Under the conditions of the theorem, using (24), we have

$$\boldsymbol{\varepsilon}'_{C_{\Omega\omega}\cap C_{0,j}}P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}}\boldsymbol{\varepsilon}_{C_{\Omega\omega}\cap C_{0,j}} = O\left(\log\log N_{\omega|j}\right) = O\left(\log\log N\right) \text{ a.s.}, \tag{41}$$

where $N_{\omega|j} = |C_{\Omega\omega} \cap C_{0,j}|$. Thus it follows that $F_N(\Pi_\Omega) - F_N(\Pi_0) > 0$ a.s. for N large enough.

B2. Underparameterised case: $\Omega < \Omega_0$.

Again we want to show that for N large enough, $F_N(\Pi_\Omega) - F_N(\Pi_0) > 0$ a.s. In this case, $(f(\Omega) - f(\Omega_0)) < 0$ and by assumption, $\lim_{N\to\infty} N^{-1}\theta_N = 0$. The result will follow if we show that $N\log(RSS_T(\Omega)/RSS_T(\Omega_0))$ is positive and of order N.

The following lemma is necessary for our proof.

Lemma 4 *Suppose that Assumption CA.2 holds true. Then, for any possible partition* Π_Ω *with* $\Omega < \Omega_0$*, there exist* $C_{\Omega\omega} \in \Pi_\Omega$ *and* $C_{0,\omega_1}, C_{0,\omega_2} \in \Pi_0$ *such that*

$$|C_{\Omega\omega} \cap C_{0,\omega_1}| > c_0 N \textit{ and } |C_{\Omega\omega} \cap C_{0,\omega_2}| > c_0 N \textit{ for any } \omega \textit{ and } N \textit{ large enough,} \tag{42}$$

where c_0 *is a fixed constant.*

Proof. See Shao and Wu (2005), Lemma 3.1. ■

From Lemma 3, for any partition $\Pi_\Omega = \{C_{\Omega 1}, ..., C_{\Omega\Omega}\}$, there exists one cluster in Π_Ω, say $C_{\Omega 1}$, and two distinct true clusters $C_{0,1}$ and $C_{0,2}$, such that

$$c_0 N < |C_{\Omega 1} \cap C_{0,1}| < N \text{ and } c_0 N < |C_{\Omega 1} \cap C_{0,2}| < N, \tag{43}$$

for N large enough. Denote the family of subsets $\{C_{\Omega\omega} \cap C_{0,j} : j = 1, ..., \Omega_0,\ \omega = 1, ..., \Omega\} - \{C_{\Omega 1} \cap C_{0,1},\ C_{\Omega 1} \cap C_{0,2}\}$ by $\mathcal{L}_{\overline{12}}$. Then

$$\begin{aligned}
& RSS_T(\Omega) - RSS_T(\Omega_0) \\
= \ & T^{-1}\left(\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{X}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{X}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right) \\
= \ & T^{-1}\left(\left\|\widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,1}} - \widetilde{X}_{C_{\Omega 1}\cap C_{0,1}}\widehat{\boldsymbol{\beta}}_{\Omega 1}\right\|^2 + \left\|\widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,2}} - \widetilde{X}_{C_{\Omega 1}\cap C_{0,2}}\widehat{\boldsymbol{\beta}}_{\Omega 1}\right\|^2\right) + \\
& T^{-1}\left(\sum_{\mathcal{L}_{\overline{12}}}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_{0,j}} - \widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{X}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right).
\end{aligned} \tag{44}$$

Let $\widetilde{X}_{11} = \widetilde{X}_{C_{\Omega 1}\cap C_{0,1}}$, $\widetilde{X}_{11a} = \left(\widetilde{X}_{11}'\ \mathbf{0}_{K\times|C_{\Omega 1}\cap C_{0,2}|}\right)'$, $\widetilde{X}_{12} = \widetilde{X}_{C_{\Omega 1}\cap C_{0,2}}$,

$$\widetilde{\mathbf{Y}}_{012} = \begin{pmatrix} \widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,1}} \\ \widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,2}} \end{pmatrix},\ \widetilde{X}_{012} = \begin{pmatrix} \widetilde{X}_{11} \\ \widetilde{X}_{12} \end{pmatrix},\ \widetilde{\boldsymbol{\varepsilon}}_{012} = \begin{pmatrix} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega 1}\cap C_{0,1}} \\ \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega 1}\cap C_{0,2}} \end{pmatrix}. \tag{45}$$

Hence

$$\begin{aligned}
& RSS_T(\Omega) - RSS_T(\Omega_0) \\
\geq \ & T^{-1}\left(\left\|\widetilde{\mathbf{Y}}_{012} - \widetilde{X}_{012}\widehat{\boldsymbol{\beta}}_{012}\right\|^2 + \sum_{\mathcal{L}_{\overline{12}}}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_{0,j}} - \widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}\widehat{\boldsymbol{\beta}}_{\omega|j}\right\|^2\right) \\
& -T^{-1}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}'\left(I - P_{\widetilde{X}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}},
\end{aligned} \tag{46}$$

where $\widehat{\boldsymbol{\beta}}_{012}$ is the least squares estimate of $\boldsymbol{\beta}$ based on $\left(\widetilde{\mathbf{Y}}_{012}, \widetilde{X}_{012}\right)$. Since $\widetilde{\mathbf{Y}}_{012} = \widetilde{X}_{012}\boldsymbol{\beta}_{02} + \widetilde{X}_{11a}\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right) + \widetilde{\boldsymbol{\varepsilon}}_{012}$, we have that

$$\begin{aligned}
& RSS_T(\Omega) - RSS_T(\Omega_0) \\
\geq \ & T^{-1}\left(\widetilde{\mathbf{Y}}_{012}'\left(I - P_{\widetilde{X}_{012}}\right)\widetilde{\mathbf{Y}}_{012} + \sum_{\mathcal{L}_{\overline{12}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}'\left(I - P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}\right) \\
& -T^{-1}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}'\left(I - P_{\widetilde{X}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} \\
= \ & T^{-1}\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right)'\widetilde{X}_{11}'\left[I - \widetilde{X}_{11}\left(\widetilde{X}_{11}'\widetilde{X}_{11} + \widetilde{X}_{12}'\widetilde{X}_{12}\right)^{-1}\widetilde{X}_{11}'\right]\widetilde{X}_{11}\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right) \\
& +T^{-1}2\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right)'\widetilde{X}_{11a}'\left(I - P_{\widetilde{X}_{012}}\right)\widetilde{\boldsymbol{\varepsilon}}_{012} + T^{-1}\widetilde{\boldsymbol{\varepsilon}}_{012}'\left(I - P_{\widetilde{X}_{012}}\right)\widetilde{\boldsymbol{\varepsilon}}_{012} + \\
& T^{-1}\sum_{\mathcal{L}_{\overline{12}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}'\left(I - P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}} - T^{-1}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}'\left(I - P_{\widetilde{X}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} \\
= \ & T^{-1}\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right)'\left[\left(\widetilde{X}_{11}'\widetilde{X}_{11}\right)^{-1} + \left(\widetilde{X}_{12}'\widetilde{X}_{12}\right)^{-1}\right]^{-1}\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right) + \\
& T^{-1}2\left(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right)'\widetilde{X}_{11a}'\left(I - P_{\widetilde{X}_{012}}\right)\widetilde{\boldsymbol{\varepsilon}}_{012} - T^{-1}\widetilde{\boldsymbol{\varepsilon}}_{012}'P_{\widetilde{X}_{012}}\widetilde{\boldsymbol{\varepsilon}}_{012} \\
& -T^{-1}\sum_{\mathcal{L}_{\overline{12}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}'P_{\widetilde{X}_{C_{\Omega\omega}\cap C_{0,j}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}} + T^{-1}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}'P_{\widetilde{X}_{C_{0,j}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}},
\end{aligned} \tag{47}$$

using the algebraic identity $(A+B)^{-1} = A^{-1} - A^{-1}\left(A^{-1}+B^{-1}\right)^{-1}A^{-1}$, where A and B are non-singular matrices.

Since we have assumed $|\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}| > 0$, given Assumption CA.2 we have

$$(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02})' \left[\left(\widetilde{X}_{11}'\widetilde{X}_{11}\right)^{-1} + \left(\widetilde{X}_{12}'\widetilde{X}_{12}\right)^{-1}\right]^{-1} (\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}) \geq c_0 N \left|\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}\right|.$$

Using (22), (24) and the Cauchy-Schwartz inequality we see that the other terms in the above lower bound are of smaller order in N. As $RSS_T(\Omega_0)/N$ is bounded away from 0 and ∞ almost surely, we have, for N large enough,

$$N \log\left(1 + \frac{RSS_T(\Omega) - RSS_T(\Omega_0)}{RSS_T(\Omega_0)}\right) > N\log(1+K),$$

for some positive K, and the result follows.

Table A1. Simulation results for $\Omega_0 = 1$, $\varepsilon_{\omega it}$ is purely idiosyncratic.

	$K = 1$				$K = 4$			
	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$
	$\zeta = 4$		$\zeta = 8$		$\zeta = 4$		$\zeta = 8$	
MIC_1								
$\Omega = 1$	**.013**	**.000**	**.013**	**.000**	**.000**	**.000**	**.000**	**.000**
$\Omega = 2$	.934	.021	.934	.021	.000	.000	.000	.000
$\Omega = 3$	.053	.979	.053	.979	1.00	1.00	1.00	1.00
MIC_2								
$\Omega = 1$	**.013**	**.018**	**.013**	**.018**	**.631**	**1.00**	**.631**	**1.00**
$\Omega = 2$	.987	.964	.987	.964	.369	.000	.369	.000
$\Omega = 3$	.000	.018	.000	.018	.000	.000	.000	.000
MIC_3								
$\Omega = 1$	**1.00**	**1.00**	**1.00**	**1.00**	**.973**	**1.00**	**.973**	**1.00**
$\Omega = 2$	.000	.000	.000	.000	.027	.000	.027	.000
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega = 1$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000

Table A2. Simulation results for $\Omega_0 = 2$, $\varepsilon_{\omega it}$ is purely idiosyncratic.

	$K = 1$				$K = 4$			
	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$
	$\zeta = 4$		$\zeta = 8$		$\zeta = 4$		$\zeta = 8$	
MIC_1								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**.009**	**.000**	**.202**	**.000**	**.188**	**.000**	**.210**	**.000**
$\Omega = 3$	.241	.000	.384	.052	.312	.061	.381	.033
$\Omega = 4$	.750	1.00	.414	.948	.500	.939	.409	.967
MIC_2								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**.157**	**.451**	**.210**	**.482**	**.183**	**.291**	**.410**	**.476**
$\Omega = 3$	.343	.482	.389	.495	.395	.307	.403	.419
$\Omega = 4$	.500	.067	.401	.023	.422	.402	.187	.105
MIC_3								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000

Table A3. Simulation results for $\Omega_0 = 3$, $\varepsilon_{\omega it}$ is purely idiosyncratic.

	$K = 1$				$K = 4$			
	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$
	$\zeta = 4$		$\zeta = 8$		$\zeta = 4$		$\zeta = 8$	
MIC_1								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.028
$\Omega = 3$	**.528**	**.577**	**.616**	**.629**	**.023**	**.000**	**.024**	**.000**
$\Omega = 4$	.472	.421	.384	.371	.432	.000	.441	.000
$\Omega = 5$	.000	.002	.000	.000	.545	1.00	.535	1.00
MIC_2								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 3$	**.633**	**.715**	**.679**	**.749**	**.841**	**.142**	**.890**	**.237**
$\Omega = 4$	.367	.285	.321	.251	.159	.545	.110	.454
$\Omega = 5$	.000	.000	.000	.000	.000	.313	.000	.309
MIC_3								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 3$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 5$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.103	.006	.000	.000
$\Omega = 3$	**1.00**	**1.00**	**1.00**	**1.00**	**.897**	**.994**	**1.00**	**1.00**
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 5$	.000	.000	.000	.000	.000	.000	.000	.000

Table A1(b). Simulation results for $\Omega_0 = 1$, $\varepsilon_{\omega it} = \lambda_{\omega i}\phi_t + \upsilon_{\omega it}$.

	$K = 1$				$K = 4$			
	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$
	$\zeta = 4$		$\zeta = 8$		$\zeta = 4$		$\zeta = 8$	
MIC_1								
$\Omega = 1$	**.015**	**.000**	**.015**	**.000**	**.000**	**.000**	**.000**	**.000**
$\Omega = 2$	.934	.023	.934	.023	.000	.000	.000	.000
$\Omega = 3$	.051	.977	.051	.977	1.00	1.00	1.00	1.00
MIC_2								
$\Omega = 1$	**.021**	**.032**	**.021**	**.032**	**.533**	**.698**	**.533**	**.698**
$\Omega = 2$	.979	.968	.979	.968	.369	.302	.369	.302
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_3								
$\Omega = 1$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 2$	.000	.000	.000	.000	.027	.000	.027	.000
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega = 1$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000

Table A2(b). Simulation results for $\Omega_0 = 2$, $\varepsilon_{\omega it} = \lambda_{\omega i}\phi_t + \upsilon_{\omega it}$.

	$K = 1$				$K = 4$			
	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$
	$\zeta = 4$		$\zeta = 8$		$\zeta = 4$		$\zeta = 8$	
MIC_1								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**.109**	**.000**	**.231**	**.000**	**.172**	**.000**	**.210**	**.000**
$\Omega = 3$	.281	.000	.401	.069	.222	.011	.381	.033
$\Omega = 4$	.610	1.00	.368	.931	.606	.989	.409	.967
MIC_2								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**.196**	**.299**	**.279**	**.412**	**.258**	**.323**	**.396**	**.431**
$\Omega = 3$	.383	.589	.332	.587	.495	.376	.501	.569
$\Omega = 4$	.421	.112	.396	.001	.247	.301	.103	.000
MIC_3								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega = 1$	.069	.009	.000	.000	.086	.000	.000	.000
$\Omega = 2$	**.931**	**.991**	**1.00**	**1.00**	**.914**	**1.00**	**1.00**	**1.00**
$\Omega = 3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000

Table A3(b). Simulation results for $\Omega_0 = 3$, $\varepsilon_{\omega it} = \lambda_{\omega i}\phi_t + \upsilon_{\omega it}$.

	$K = 1$				$K = 4$			
	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$	$N = 100$	$N = 400$
	$\zeta = 4$		$\zeta = 8$		$\zeta = 4$		$\zeta = 8$	
MIC_1								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.028
$\Omega = 3$	**.541**	**.496**	**.603**	**.568**	**.026**	**.000**	**.049**	**.000**
$\Omega = 4$	.459	.504	.397	.432	.439	.000	.491	.000
$\Omega = 5$	.000	.002	.000	.000	.535	1.00	.460	1.00
MIC_2								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 3$	**.671**	**.727**	**.715**	**.768**	**.811**	**.310**	**.863**	**.391**
$\Omega = 4$	.329	.273	.285	.232	.189	.479	.137	.491
$\Omega = 5$	.000	.000	.000	.000	.000	.211	.000	.118
MIC_3								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 3$	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 5$	.000	.000	.000	.000	.000	.000	.000	.000
MIC_4								
$\Omega = 1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 2$	.088	.027	.000	.000	.151	.049	.000	.000
$\Omega = 3$	**.912**	**.973**	**1.00**	**1.00**	**.849**	**.951**	**1.00**	**1.00**
$\Omega = 4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega = 5$	.000	.000	.000	.000	.000	.000	.000	.000

Table A4. Finite sample properties of estimators, $\varepsilon_{\omega it}$ is purely idiosyncratic.

	$K=1$, $\Omega_0=2$, $\overline{\beta}=0.85$, $\boldsymbol{\beta}=(1,.5)'$				$K=1$, $\Omega_0=3$, $\overline{\beta}=0.475$, $\boldsymbol{\beta}=(1,.5,-.25)'$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
pooled FE	.688	.678	.688	.678	-.027	-.039	-.026	-.042
	(.013)	(.007)	(.009)	(.005)	(.006)	(.004)	(.004)	(.002)
pooled OLS	672	.680	.675	.679	-.047	-.038	-.045	-.036
	(.018)	(.006)	(.013)	(.004)	(.009)	(.004)	(.008)	(.003)
FE_1	1.02	1.02	1.02	1.01	1.03	1.03	1.02	1.01
	(.057)	(.009)	(.016)	(.007)	(.026)	(.013)	(.019)	(.010)
FE_2	.512	.506	.502	.502	.506	.503	.502	.501
	(.056)	(.008)	(.010)	(.006)	(.017)	(.009)	(.012)	(.007)
FE_3					-.251	-.250	-.250	-.250
					(.006)	(.005)	(.004)	(.003)

Table A4(b). Finite sample properties of estimators, $\varepsilon_{\omega it} = \lambda_{\omega i}\phi_t + \upsilon_{\omega it}$.

	$K=1$, $\Omega_0=2$, $\overline{\beta}=0.85$, $\boldsymbol{\beta}=(1,.5)'$				$K=1$, $\Omega_0=3$, $\overline{\beta}=0.475$, $\boldsymbol{\beta}=(1,.5,\text{-}.25)'$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
pooled FE	.684	.671	.685	.677	-.026	-.031	-.024	-.037
	(.013)	(.008)	(.009)	(.005)	(.008)	(.005)	(.005)	(.003)
pooled OLS	683	.670	.689	.681	-.025	-.033	-.024	-.037
	(.014)	(.009)	(.010)	(.006)	(.009)	(.005)	(.007)	(.004)
FE_1	1.03	1.01	1.02	1.01	1.03	1.02	1.01	1.00
	(.028)	(.018)	(.016)	(.009)	(.034)	(.017)	(.022)	(.009)
FE_2	.505	.502	.502	.501	.501	.501	.501	.502
	(.019)	(.011)	(.010)	(.007)	(.020)	(.010)	(.014)	(.008)
FE_3					-.247	-.251	-.251	-.250
					(.009)	(.005)	(.005)	(.004)