To Pool or Not To Pool: A Partially Heterogeneous Alternative

Vasilis Sarafidis* University of Sydney

Neville Weber† University of Sydney

This version: January 2009

Abstract

This paper proposes a new modeling framework for the analysis of panel data based on the concept of 'partitional clustering'. In particular, the population of cross-sections is grouped into clusters, such that parameter homogeneity is maintained only within clusters. To determine the (unknown) number of clusters we put forward an information-based criterion, which, as we show, is strongly consistent for fixed T − in other words, it selects the correct number of clusters with probability 1 as the number of cross-sections grows large. Simulation experiments show that the proposed criterion performs well even with moderately small N. We apply the method in a panel data set of commercial banks and we find three clusters, with significant differences in the slope parameters across them.

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

JEL Classification: C13; C33; C51.

1 Introduction

Full homogeneity in the slope parameters of a panel data model is very 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 letting these

*Corresponding author. Discipline of Econometrics and Business Statistics, 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.

parameters be fully heterogeneous 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. 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 a relatively large T, heterogeneous models for individual states tend to produce implausible estimates with inferior forecasting properties despite 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 two polar cases, and other intermediate cases may 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 that are 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 cross-sections for reasons attributed to budget constraints, arbitrage conditions and common technologies. This procedure is appealling because it imposes constraints that are directly related to economic theory, however, it is applicable only in panels with a large number of cross-sectional and time series observations − hereafter, denoted by N and T, respectively. As Pesaran, Shin and Smith point out, for small T PMGE is likely to be severely biased.

This paper proposes a modelling framework that imposes partially heteroge-

neous restrictions not with respect to time, as PMGE does, but with respect to the cross-sectional dimension of the panel. In particular, the population of cross-sections is grouped into different clusters, such that within each cluster the parameters are homogeneous and all inter-cluster heterogeneity is attributed to the usual individual-specific time-invariant effects. The clusters themselves are heterogeneous − that is, the regression coefficients vary across clusters.

Naturally, the practical issue of how to cluster the cross-sections into homogeneous groups is central in this paper. Clustering methods have already been advocated in the econometric panel data literature by a number of researchers, such as in Durlauf and Johnson (1995), who propose clustering the individuals using regression tree analysis; Vahid (1999), who puts forward a classification algorithm based on a measure of complexity employed in coding theory, using the principles of minimum description length and minimum message length; and Kapetanios (2006), who proposes an information criterion method, based on simulated annealing. However, all these methods are based on the concept of *hierarchical clustering*, which involves building a 'hierarchy' from the individual cross-sections by progressively merging them into larger clusters. As a result, the proposed procedures are appropriate only when T is sufficiently large − otherwise, the estimated parameters can have very poor properties in terms of classifying the individuals into the appropriate clusters. For this reason, 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, the clustering method proposed in this paper relies on the concept of *partitional clustering*. The idea is simple; instead of treating each individual as a distinct cluster to begin with (as in hierarchical clustering) which requires a large T available, 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 the residual sum of squares (RSS) of the estimated model. The same procedure is followed for different numbers of clusters with the aim of finding the optimal number of clusters, i.e. 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 there-

fore it tends to over-parameterise the model by allowing for more clusters than they actually exist. Hence, the penalty acts essentially as a filter to ensure that the preferred clustering outcome partitions between clusters rather than within clusters. We show in the paper that this procedure is fixed-T consistent, in the sense that the probability of estimating the correct number of clusters approaches 1 as $N \to \infty$. The intuition of this method is very similar to that of a model selection criterion for choosing the number of parameters in a regression model, although the study of the asymptotics is far more complicated in this case because the number of individuals contained in a given cluster may vary with N.

Once the number of clusters and membership in these clusters are determined, the classic homogeneity assumption imposed in standard panel data models is then testable. The suggested procedure for dealing with parameter heterogeneity is fairly simple. Therefore, it provides an appealing solution to the problem of estimating heterogeneous panels – particularly when dealing with applications that have a large number of cross-sections and a small, or only moderately large, number of time-series observations. Financial and microeconomic data sets are straightforward examples of the panels we have in mind.

The remainder of the paper is as follows. The next section sets out the partially heterogeneous model proposed in this paper and the assumptions we make. 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 partitional clustering algorithm. The performance of the algorithm is investigated in Section 5 using Monte Carlo experiments. Section 6 illustrates the technique using a balanced panel of 1,350 banking institutions operating in the U.S., each observed over a period of 15 years and this is followed by a brief concluding section.

2 Model Specification

We consider the following panel data model:

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

where $y_{\omega it}$ denotes the observation on the dependent variable for the ith 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, $\alpha_{\omega i}$ denotes an individual-specific, time-invariant unobserved effect 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 number of observations is given by $S = NT$.

Essentially, the model given in (1) makes a case for pooling the data within clusters of cross-sections and allowing for slope parameter heterogeneity across clusters. One way to rationalise this is on the basis of the existence of certain factors or qualities, which differentiate the population of cross-sections into groups, such that within each group the cross-sections respond similarly to changes in the conditional regressors and all inter-cluster heterogeneity is captured by individual-specific effects. On the other hand, cross-sections that belong to different clusters may vary in terms of these factors/qualities and therefore they respond in a different manner. One example where this set up may apply in practice is a firm data set, where some plausible qualities that differentiate the individual firms into clusters of cross-sections may be the type of industry within which they operate, the size of the firm, the type of ownership and so on.[1]

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

BA.1 $E(\alpha_{\omega i}) = 0$, $E(\alpha_{\omega i}^2) = \sigma_{\alpha_{\omega i}}^2 < \infty$.

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

BA.3 $N_\omega^{-1} \sum_{i=1}^{N_\omega} \mathbf{X}_{\omega i}' \mathbf{X}_{\omega i} \xrightarrow{p} \mathbf{Q}_{XX,\omega}$, finite and positive definite.

BA.4 There exists a fixed constant, $0 < c_\omega < 1$, such that $N_\omega / N \to c_\omega$ for $\omega = 1, ..., \Omega$, as $N \to \infty$.

[1] At this point it is useful to distinguish this type of clustering from the usual notion of clustering that is common in homogeneous panel data econometrics, which refers to the idea that the errors are independent across groups but not within groups of cross-sections. In this case, it is well known that using OLS-type estimates of the standard errors can bias standard errors downwards (assuming that the intra-cluster correlations are positive), leading to over-rejecting the null hypothesis that the regression coefficients are zero. The model proposed here is different in the sense that clustering refers to differences in the parameter values of the regression across groups of cross-sections rather than to existing correlations in the error covariance matrix within groups.

BA.5 Ω is a fixed unknown constant, such that $0 < \Omega \leq \xi$, where ξ is fixed and known.

BA.6 $\boldsymbol{\beta}_\omega = \boldsymbol{\beta} + \boldsymbol{\eta}_\omega$, where $\boldsymbol{\eta}_\omega$ is a $K \times 1$ vector of fixed constants, such that $\sum_{\omega=1}^{\Omega} c_\omega \boldsymbol{\eta}_\omega = \mathbf{0}$.

Assumptions BA.1–BA.2 reflect the familiar error structure in a panel model with strictly exogenous regressors. Assumption BA.3 ensures that $\left(N_\omega^{-1} \sum_{i=1}^{N_\omega} \mathbf{X}'_{\omega i} \mathbf{X}_{\omega i} \right)^{-1}$ exists in probability for all N_ω sufficiently large. Assumption BA.4 ensures that no clusters are asymptotically negligible. The asymptotics can be conceived via 'class-growing sequences', as in Shao and Wu (2005). Assumption BA.5 ensures that the total number of clusters is bounded by a constant ξ. Assumption BA.6 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.

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

3 On the Impact of Partial Heterogeneity

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

$$\mathbf{y}_{\omega i} = \mathbf{X}_{\omega i} \boldsymbol{\beta}_\omega + \mathbf{e}_T \alpha_{\omega i} + \boldsymbol{\varepsilon}_{\omega i}, \tag{2}$$

where $\mathbf{y}_{\omega i} = (y_{\omega i1}, ..., y_{\omega iT})'$, $\mathbf{X}_{\omega i} = (\mathbf{x}_{\omega i1}, ..., \mathbf{x}_{\omega iT})'$, $\boldsymbol{\varepsilon}_{\omega i} = (\varepsilon_{\omega i1}, ..., \varepsilon_{\omega iT})'$ and $\mathbf{e}_T$ is a $T \times 1$ vector of ones.

Ignoring the partially heterogeneous structure in (2) results in the following regression model:

$$\mathbf{y}_{\omega i} = \mathbf{X}_{\omega i} \boldsymbol{\beta} + \mathbf{e_T} \alpha_{\omega i} + \boldsymbol{\upsilon}_{\omega i}, \; \boldsymbol{\upsilon}_{\omega i} = \boldsymbol{\varepsilon}_{\omega i} + \mathbf{X}_{\omega i} \boldsymbol{\eta}_\omega. \tag{3}$$

Define the $T \times T$ idempotent matrix $\mathbf{Q}_T = \mathbf{I}_T - T^{-1} \mathbf{e}_T \mathbf{e}'_T$, which transforms the observations in terms of deviations from individual-specific averages and sweeps out the individual effects, $\alpha_{\omega i}$. We have

$$\mathbf{Q}_T \mathbf{y}_{\omega i} = \mathbf{Q}_T \mathbf{X}_{\omega i} \boldsymbol{\beta} + \mathbf{Q}_T \boldsymbol{\upsilon}_{\omega i}, \; \mathbf{Q}_T \boldsymbol{\upsilon}_{\omega i} = \mathbf{Q}_T \boldsymbol{\varepsilon}_{\omega i} + \mathbf{Q}_T \mathbf{X}_{\omega i} \boldsymbol{\eta}_\omega, \tag{4}$$

or

$$\widetilde{\mathbf{y}}_{\omega i} = \widetilde{\mathbf{X}}_{\omega i}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}_{\omega i} + \widetilde{\mathbf{X}}_{\omega i}\boldsymbol{\eta}_{\omega}, \tag{5}$$

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

The fixed effects estimator is given by:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{FE} &= \left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega} \widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{X}}_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega} \widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{y}}_{\omega i}\right] \\
&= \boldsymbol{\beta} + \left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega} \widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{X}}_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega} \left(\widetilde{\mathbf{X}}'_{\omega i}\widetilde{\boldsymbol{\varepsilon}}_{\omega i} + \widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{X}}_{\omega i}\boldsymbol{\eta}_{\omega}\right)\right].
\end{aligned} \tag{6}$$

Taking plims over N yields:

$$\begin{aligned}
&\text{plim}_{N\to\infty}\left(\widehat{\boldsymbol{\beta}}_{FE} - \boldsymbol{\beta}\right) = \\
= \ &\text{plim}_{N\to\infty}\left[\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{X}}_{\omega i}\right)\right]^{-1}\left\{\text{plim}_{N\to\infty}\left[\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{\mathbf{X}}'_{\omega i}\widetilde{\boldsymbol{\varepsilon}}_{\omega i}\right)\right]\right. \\
&\left.+\text{plim}_{N\to\infty}\left[\sum_{\omega=1}^{\Omega}\frac{N_\omega}{N}\left(\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{X}}_{\omega i}\right)\boldsymbol{\eta}_{\omega}\right]\right\} \\
= \ &\left[\sum_{\omega=1}^{\Omega} c_\omega \widetilde{\mathbf{Q}}_{XX,\omega}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\widetilde{\mathbf{Q}}_{XX,\omega}c_\omega\boldsymbol{\eta}_{\omega}\right],
\end{aligned} \tag{7}$$

where $\widetilde{\mathbf{Q}}_{XX,\omega} = \text{plim}_{N_\omega\to\infty}\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\widetilde{\mathbf{X}}'_{\omega i}\widetilde{\mathbf{X}}_{\omega i}$ and its existence is guaranteed given BA.3. As we can see, the fixed effects estimator is not necessarily consistent. In particular, consistency is only achieved in this case when $\widetilde{\mathbf{Q}}_{XX,\omega}c_\omega\boldsymbol{\eta}_{\omega}$ sums up to zero and this will happen, for example, when the limiting matrix $\widetilde{\mathbf{Q}}_{XX,\omega}$ is orthogonal to the vector $c_\omega\boldsymbol{\eta}_{\omega}$ for all ω. This would happen if, say, the $\widetilde{\mathbf{Q}}_{XX,\omega}$ matrices are constant across clusters. However, there is no guarantee that this condition will hold true in any empirical situation.

This result is surprising given that in a fully heterogeneous model, where each individual forms its own cluster that gives rise to the following model[2]:

$$y_{it} = \alpha_i + \boldsymbol{\beta}'_i\mathbf{x}_{it} + \varepsilon_{it} \text{ with } \boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\eta}_i, \tag{8}$$

strict exogeneity, which implies that $E\left(\boldsymbol{\eta}_i|\mathbf{x}_{i1},...,\mathbf{x}_{iT}\right) = 0$, is sufficient to ensure consistency of $\widehat{\boldsymbol{\beta}}_{FE}$, while in a fully homogeneous model $\widehat{\boldsymbol{\beta}}_{FE}$ is also consistent of course under the basic assumptions of the model because $\boldsymbol{\eta}_{\omega} = 0$. However, as we

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

see here in the intermediate case of partial heterogeneity, $\widehat{\boldsymbol{\beta}}_{FE}$ does not converge to $\boldsymbol{\beta}$ in general.

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 $\alpha_{\omega i}$ terms are uncorrelated with $E\left(\alpha_{\omega it}|\mathbf{x}_{\omega i1},...,\mathbf{x}_{\omega iT}\right)=0\ \forall\ \omega$ and i.

The estimable model is now given by:

$$y_{\omega it}=\boldsymbol{\beta}'\mathbf{x}_{\omega it}+\left(\varepsilon_{\omega it}+\alpha_{\omega i}+\boldsymbol{\eta}_{\omega}'\mathbf{x}_{\omega it}\right), \tag{9}$$

or in matrix form:

$$\mathbf{y}_{\omega i}=\mathbf{X}_{\omega i}\boldsymbol{\beta}+\mathbf{v}_{\omega i},\ \mathbf{v}_{\omega i}=\boldsymbol{\varepsilon}_{\omega i}+\mathbf{e}_T\alpha_{\omega i}+\mathbf{X}_{\omega i}\boldsymbol{\eta}_{\omega}.$$

Hence, the pooled OLS estimator for $\boldsymbol{\beta}$ is given in this case by:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{OLS} &= \left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\mathbf{X}_{\omega i}'\mathbf{X}_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\mathbf{X}_{\omega i}'\mathbf{y}_{\omega i}\right]\\ &= \boldsymbol{\beta}+\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\mathbf{X}_{\omega i}'\mathbf{X}_{\omega i}\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\sum_{i=1}^{N_\omega}\mathbf{X}_{\omega i}'\left(\boldsymbol{\varepsilon}_{\omega i}+\mathbf{e}_T\alpha_{\omega i}+\mathbf{X}_{\omega i}\boldsymbol{\eta}_{\omega}\right)\right]. \end{aligned}\tag{10}$$

Taking plims over N yields[3]:

$$\begin{aligned}&\text{plim}_{N\to\infty}\left(\widehat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{\beta}\right)=\\ &= \left[\sum_{\omega=1}^{\Omega}\left(\widetilde{\mathbf{Q}}_{XX,\omega}+T\widetilde{\mathbf{M}}_{XX,\omega}\right)c_\omega\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\left(\widetilde{\mathbf{Q}}_{XX,\omega}+T\widetilde{\mathbf{M}}_{XX,\omega}\right)c_\omega\boldsymbol{\eta}_{\omega}\right], \end{aligned}\tag{11}$$

where $\widetilde{\mathbf{M}}_{XX,\omega}=\text{plim}_{N_\omega\to\infty}\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\overline{\mathbf{X}}_{\omega i}'\overline{\mathbf{X}}_{\omega i}$ with $\overline{\mathbf{X}}_{\omega i}=T^{-1}e_T'\mathbf{X}_{\omega i}$, while $\widetilde{\mathbf{Q}}_{XX,\omega}$ has been defined below (7). Hence, from (7) and (11) we can see that the difference between the two estimators is that consistency for $\widehat{\boldsymbol{\beta}}_{OLS}$ is guaranteed only when both $\widetilde{\mathbf{Q}}_{XX,\omega}$ and $\widetilde{\mathbf{M}}_{XX,\omega}$ are constant across clusters, while $\widehat{\boldsymbol{\beta}}_{FE}$ requires only $\widetilde{\mathbf{Q}}_{XX,\omega}$ to be constant since the observations are transformed in terms of deviations from individual-specific averages.

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, $\Pi_0=\{C_{0,1},\ldots,C_{0,\Omega_0}\}$

[3] See Appendix A.

with $C_{0,\omega}$ being the set of indices of elements in the ωth cluster, $C_{0,\omega} = \{\omega_1, \ldots, \omega_{N_{0\omega}}\} \subseteq \{1, 2, \ldots, N\}$. The number of individuals in the ωth cluster is $|C_{0,\omega}| = N_{0\omega}$, $N_{01} + \ldots + N_{0\Omega_0} = N$.

Each cluster has its own regression structure so

$$\mathbf{Y}_{C_{0,\omega}} = \mathbf{X}_{C_{0,\omega}}\boldsymbol{\beta}_{0\omega} + \alpha_{C_{0,\omega}} + \varepsilon_{C_{0,\omega}}, \text{ for } \omega = 1, ..., \Omega_0, \tag{12}$$

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 vector of observations on y for the individuals in the ωth cluster and $\alpha_{C_{0,\omega}} = \left(\mathbf{e}'_T\alpha_{\omega_1}, ..., \mathbf{e}'_T\alpha_{\omega_{N_{0\omega}}}\right)'$. The $\boldsymbol{\beta}_{0\omega}$ are vectors of fixed coefficients. We assume the $\boldsymbol{\beta}_{0\omega}$ vectors are distinct. The $\mathbf{X}_{C_{0,\omega}}$ matrix is the corresponding $(N_{0\omega}T) \times K$ matrix of covariates.

Premultiplying (12) by $\mathbf{Q}_{C_{0,\omega}} = \mathbf{I}_{N_{0\omega}T} - \mathbf{I}_{N_{0\omega}} \otimes \frac{1}{T}\mathbf{e}_T\mathbf{e}'_T$ to remove the individual-specific effects yields:

$$\mathbf{Q}_{C_{0,\omega}}\mathbf{Y}_{C_{0,\omega}} = \mathbf{Q}_{C_{0,\omega}}\mathbf{X}_{C_{0,\omega}}\boldsymbol{\beta}_{0\omega} + \mathbf{Q}_{C_{0,\omega}}\boldsymbol{\varepsilon}_{C_{0,\omega}}, \tag{13}$$

where $\mathbf{I}_k$ is a $k \times k$ identity matrix, or

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

where $\widetilde{\mathbf{Y}}_{C_{0,\omega}} = \mathbf{Q}_{C_{0,\omega}}\mathbf{Y}_{C_{0,\omega}}$, 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 situation 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.4 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.2 to

CA.1 Given the covariates $\{x_{\omega it}\}$ corresponding to the observations in the ωth cluster, the error vectors $\{\varepsilon_{\omega it}\} = (\varepsilon_{\omega i1}, ..., \varepsilon_{\omega iT})'$ for the individuals in the cluster are independent and identically distributed random vectors with mean vector $\mathbf{0}$, positive-definite covariance matrix $\boldsymbol{\Sigma}_\omega$, and for some $\delta > 0$, $E\left|\varepsilon_{\omega it}\right|^{2+\delta} < \infty$.

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

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

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

$$\sum_{i=1}^{N_\ell T} \left| x_{\omega\ell}^{(i)} \right|^{2+\delta} = O\left[\left(\mathbf{x}'_{\omega\ell}\mathbf{x}_{\omega\ell} \right)^{(2+\delta)/2} / \log\left(\mathbf{x}'_{\omega\ell}\mathbf{x}_{\omega\ell} \right)^{1+\delta} \right] \tag{15}$$

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 $\mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}}$ denote the projection matrix

$$\mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}} = \widetilde{\mathbf{X}}_{C_\ell} \left(\widetilde{\mathbf{X}}'_{C_\ell} \widetilde{\mathbf{X}}_{C_\ell} \right)^{-1} \widetilde{\mathbf{X}}'_{C_\ell}, \tag{16}$$

based on the corresponding $\widetilde{\mathbf{X}}_{C_\ell}$ matrix. Let $\boldsymbol{\varepsilon}_{C_\ell}$ denote the vector of corresponding error terms.

Lemma 1 *Let $\varpi_1, \varpi_2, ...$ be a sequence of random variables with zero mean, such that $0 < E(\varpi_i^2) = \sigma_i^2$ and $E|\varpi_i|^{2+\delta} < \tau < \infty$ 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*

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

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

$$\boldsymbol{\varepsilon}_{C_\ell} = \boldsymbol{\nu}_{C_\ell}^{(1)} + ... + \boldsymbol{\nu}_{C_\ell}^{(T)}, \tag{17}$$

where the ith element of $\boldsymbol{\nu}_{C_\ell}^{(t)}$ is $(\boldsymbol{\varepsilon}_{C_\ell})_i I(i \in \{t, t+T, T+2T, ...\})$. For example, $\boldsymbol{\nu}_{C_\ell}^{(1)} = (\varepsilon_{\omega 11}, 0, ..., 0, \varepsilon_{\omega 21}, 0, ..., \varepsilon_{\omega N_\ell 1}, 0, ..., 0)'$. 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

$$\boldsymbol{\varepsilon}'_{C_\ell} P_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = \sum_{t=1}^{T} \sum_{s=1}^{T} \boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}. \tag{18}$$

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

$$\begin{aligned}\left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)^2 &= \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} \mathbf{P}^2_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)^2 \\ &= \left(\left(\mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)}\right)' \left(\mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right)\right)^2 \\ &\leq \left(\boldsymbol{\nu}_{C_\ell}^{(t)\prime} \mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)}\right)\left(\boldsymbol{\nu}_{C_\ell}^{(s)\prime} \mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(s)}\right). \end{aligned} \tag{19}$$

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

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

$$\nu_{C_\ell}^{(t)\prime} \widetilde{\mathbf{X}}_{C_\ell} = O\left(N_\ell \log\log N_\ell\right)^{\frac{1}{2}}. \tag{20}$$

Hence,

$$\boldsymbol{\varepsilon}'_{C_\ell} \widetilde{\mathbf{X}}_{C_\ell} = O\left(\left(N_\ell \log\log N_\ell\right)^{\frac{1}{2}}\right) \text{ a.s.} \tag{21}$$

Furthermore, the assumption CA.2 ensures that the elements of $\left(\widetilde{\mathbf{X}}'_{C_\ell} \widetilde{\mathbf{X}}'_{C_\ell}\right)^{-1}$ are $O\left(N_\ell^{-1}\right)$. Therefore, using (20) and arguing as in the proof of Lemma A.2 of Bai et al. (1999) we have that

$$\begin{aligned}\boldsymbol{\nu}_{C_\ell}^{(t)\prime} P_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\nu}_{C_\ell}^{(t)} &= \boldsymbol{\nu}_{C_\ell}^{(t)\prime} \widetilde{\mathbf{X}}_{C_\ell} \left(\widetilde{\mathbf{X}}'_{C_\ell} \widetilde{\mathbf{X}}_{C_\ell}\right)^{-1} \widetilde{\mathbf{X}}'_{C_\ell} \boldsymbol{\nu}_{C_\ell}^{(t)} \\ &= O\left(\log\log N_\ell\right). \end{aligned} \tag{22}$$

As a result,

$$\boldsymbol{\varepsilon}'_{C_\ell} \mathbf{P}_{\widetilde{\mathbf{X}}_{C_\ell}} \boldsymbol{\varepsilon}_{C_\ell} = O\left(\log\log N_\ell\right). \tag{23}$$

The results in (21) and (23) 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.

$$C_{0,\omega}^{(N)} \subseteq C_{0,\omega}^{(N+1)} \text{ for all } \omega = 1, ..., \Omega_0, \text{ for large } N. \tag{24}$$

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

$C_{\Omega\omega} \cap C_{0,j}$, $\omega = 1, ..., \Omega$, $j = 1, ..., \Omega_0$. We use the following model information-based criterion (MIC) as the basis for determining the underlying cluster structure:

$$F_N\left(\Pi_{\omega}^{(N)}\right) = T^{-1}\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \psi\left(\Omega\right)\theta_N, \tag{25}$$

where $\psi\left(\Omega\right)$ is a strictly increasing function of Ω and θ_N is a sequence of constants. Thus, the first term is the residual sum of squares and represents the goodness of fit of the model, while the second term is the penalty for overfitting.

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

$$F_N\left(\Pi_{\widehat{\Omega}_0}^{(N)}\right) = \min_{1\leq\Omega\leq\xi}\min_{\Pi_{\Omega}^{(N)}} F_N\left(\Pi_{\Omega}^{(N)}\right). \tag{26}$$

The following theorem shows that the criterion in (25) 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* $\lim_{N\to\infty}\left(\log\log N\right)^{-1}\theta_N = \infty$. *Suppose that assumptions CA.1-CA.3 and BA.4 hold and* Π_0 *is the true cluster structure corresponding to model* (12). *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 1 for large* N.

Proof. See Appendix B. ■

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 study of combinatorics.[4] Stirling numbers of the second kind are given by the explicit formula:

$$S\left(N,\Omega\right) = \sum_{\omega=1}^{\Omega}\left(-1\right)^{\Omega-\omega}\frac{\omega^{N-1}}{\left(\omega-1\right)!\left(\Omega-\omega\right)!} = \frac{1}{\Omega!}\sum_{\omega=0}^{\Omega}\left(-1\right)^{\Omega-\omega}\left(\begin{array}{c}\Omega\\ \omega\end{array}\right)\omega^N. \tag{27}$$

This formula is in fact a special case of the Ω^{th} forward difference of the monomial x^N evaluated at $x = 0$:

$$\Delta^{\Omega}x^N = \sum_{\omega=0}^{\Omega}\left(-1\right)^{\Omega-\omega}\left(\begin{array}{c}\Omega\\ \omega\end{array}\right)\left(x+\omega\right)^N. \tag{28}$$

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

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

$$B_N = \sum_{\Omega=1}^{N} S(N, \Omega). \tag{29}$$

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 assume − rather optimistically − that a given computer is 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 therefore, even with small data sets − a problem that also applies in hierarchical clustering of course. Instead, 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:[5]

1. Given an initial partition and a fixed number of clusters, run the fixed effects estimator for each cluster separately and calculate the total RSS of the estimated model, which is given by $TRSS = \sum_{\omega=1}^{\Omega} RSS_\omega$;

2. Re-allocate the ith cross-section to all remaining clusters and obtain the resulting $TRSS$'s. Finally, allocate the ith individual into the cluster that achieves the smaller $TRSS$;

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

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

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

6. Pick the number of clusters that minimises

$$TRSS + \Omega\sqrt{N}. \tag{30}$$

Since only a local search over different partitions is feasible, the final outcome of partitional clustering algorithms may be sensitive to the way of initialising the data. The risk of coming across a very atypical initial partition can be

[5] The algorithm is written as an .ado file in Stata 10 and it is available to all Stata users on the web.

reduced substantially if partial heterogeneity is mainly driven by a number of observed attributes that are used to form the initial partition. Otherwise, there are methods proposed in the clustering literature to cross-validate the results (see, e.g. Kaufman and Rousseeuw, 1990).[6]

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 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 fixed effects and the pooled OLS estimators under these circumstances.

5.1 Experimental Design

The underlying process is given by:

$$\begin{aligned} y_{\omega it} &= \sum_{k=1}^{K} \beta_{\omega k} x_{k\omega it} + \alpha_{\omega i} + \varepsilon_{\omega it} = \underset{1\times K}{\boldsymbol{\beta}'_{\omega}} \underset{K\times 1}{\mathbf{x}_{\omega it}} + \alpha_{\omega i} + \varepsilon_{\omega it}, \\ \text{for } t &= 1, ..., T,\ i\,[\in \omega] = 1, ..., N_{\omega} \text{ and } \omega = 1, ..., \Omega, \end{aligned} \tag{31}$$

where $\alpha_{\omega i}$ and $\varepsilon_{\omega it}$ are drawn in each replication from $i.i.d.N\left(0, \sigma_{\alpha}^2\right)$ and $i.i.d.N\left(0, \sigma_{\varepsilon_\omega}^2\right)$ respectively, while the kth element of $\mathbf{x}_{\omega it}$ is drawn from $i.i.d.N\left(\mu_{x_\omega}, \sigma_{x_{\omega k}}^2\right)$. Let $\boldsymbol{\Sigma}_{x_\omega}$ be the $K \times K$ matrix that contains $\sigma_{x_{\omega k}}^2$ on its kth diagonal element, for $k = 1, ..., K$, and zero otherwise. Define $y^*_{\omega it} = y_{\omega it} - \alpha_{\omega i}$, such that (31) may be rewritten as

$$y^*_{\omega it} = \boldsymbol{\beta}'_{\omega} \mathbf{x}_{\omega it} + \varepsilon_{\omega it}, \tag{32}$$

and let the signal-to-noise ratio be denoted by $\zeta_\omega = \sigma_{s_\omega}^2 / \sigma_{\varepsilon_\omega}^2$, where $\sigma_{s_\omega}^2$ is the variance of the signal for the ωth cluster and $\sigma_{\varepsilon_\omega}^2$ denotes the variance of the error term. $\sigma_{s_\omega}^2$ equals

$$\sigma_{s_\omega}^2 = var\left(y^*_{\omega it} - \varepsilon_{\omega it}\right) = var\left(\boldsymbol{\beta}'_{\omega} \mathbf{x}_{\omega it}\right) = \boldsymbol{\beta}'_{\omega} \boldsymbol{\Sigma}_{x_\omega} \boldsymbol{\beta}_{\omega}. \tag{33}$$

Clearly, for a given value of $\boldsymbol{\Sigma}_{x,\omega}$ and $\sigma_{\varepsilon,\omega}^2$ the signal-to-noise ratio depends on $\boldsymbol{\beta}_\omega$. This implies that scaling the slope parameters by the same factor will produce very different results − and of course there is no natural choice for what the value

[6] One easy way to achieve this involves perturbating the observations and checking whether the clusters are robust with respect to changes in the data.

of the scaling factor should be. To control for this effect we normalise $\zeta_\omega = \zeta$ and $\sigma^2_{\varepsilon_\omega} = 1$ for $\omega = 1, ..., \Omega$, and we set the kth diagonal element of $\mathbf{\Sigma}_{x_\omega}$ equal to $\zeta / \left(\beta^2_{\omega k} K\right)$. We consider $\zeta = \{4, 8\}$, $N = \{100, 400\}$ with $T = 8$, $K = \{1, 5\}$ and $\Omega = \{1, 2, 3\}$. The values of the selected slope coefficients are listed in Table 1. 1000 replications are performed in each experiment. To reduce the computational burden, we fit models with $\omega = 1, 2, 3$ clusters when there is only one cluster, $\omega = 1, 2, 3, 4$ clusters when there are two clusters and $\omega = 1, 2, 3, 4, 5$ clusters when there are three clusters.

Table 1. Parameter values used in the simulation study.

	$K = 1$	$K = 5$
$\Omega_0 = 1$	$\beta = 1$	$\beta = \begin{pmatrix} 1 \\ .5 \\ -1.5 \\ 2 \\ .3 \end{pmatrix}$
$\Omega_0 = 2$	$\beta_1 = 1$ $\beta_2 = .5$	$\beta_1 = \begin{pmatrix} 1 \\ .5 \\ -1.5 \\ 2 \\ .3 \end{pmatrix}, \beta_2 = \begin{pmatrix} .5 \\ 1.5 \\ -2 \\ 3 \\ .7 \end{pmatrix}$
$\Omega_0 = 3$	$\beta_1 = .5$ $\beta_2 = .125$ $\beta_3 = -.25$	$\beta_1 = \begin{pmatrix} .5 \\ .8 \\ -.5 \\ 2 \\ .3 \end{pmatrix}, \beta_2 = \begin{pmatrix} .125 \\ 1.3 \\ .5 \\ 3 \\ .7 \end{pmatrix}, \beta_3 = \begin{pmatrix} -.25 \\ .2 \\ -1 \\ 1 \\ .5 \end{pmatrix}$

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. MIC denotes the model selection criterion proposed in this paper, while Ak and SIC refer to the popular Akaike and Schwartz information criteria respectively that are commonly used in variable selection. Finally, SW refers to the clustering criterion proposed by Shao and Wu (2005) in the context of a cross-sectional regression model.[7]

As we can see, MIC performs well under all circumstances. In particular, for $\Omega_0 = 1, 2$, the relative frequency of selecting the correct number of clusters using

[7]The penalty proposed by Shao and Wu is given by $A_N = \lambda^{-1}\left[(\log(N))^\lambda - 1\right]$. In their paper they experiment with different values of λ in the range between 1.0 and 8.0. Here we set $\lambda = 4$, because this one performs best in their paper.

MIC equals one. This holds true regardless of the value of N, K and ζ. For $\Omega_0 = 3$ the performance of MIC deteriorates slightly. We have explored further the underlying reason for this result. We found that when the wrong number of clusters is selected, our estimation algorithm reaches a solution that is far from the global minimum. As a result, the improvement over RSS from moving from $\Omega = 2$ to $\Omega = 3$ is only marginal and therefore the criterion errs on the side of a smaller number of clusters. Furthermore, notice that the performance of MIC naturally improves with larger values of N and ζ, while K does not appear to have a noteworthy effect − certainly a desirable property.[8]

With regards to the remaining criteria; firstly, Ak selects the largest possible number of clusters under all circumstances and therefore it seems that it is not appropriate in this context. On the other hand, the performance of SIC and SW exhibits a similar pattern in the sense that both criteria perform poorly for $\Omega_0 = 1, 2$, and they do much better for $\Omega_0 = 3$, although they still underperform compared to MIC. Specifically, we see that for $\Omega_0 = 1, 2$ the penalty imposed by SIC and SW is not large enough to avoid overfitting. Interestingly, this does not seem to carry over when $\Omega_0 = 3$, at least to the same extent. Finally, when MIC selects the wrong number of clusters it slips in the lower end because it exhibits the largest penalty for overfitting.

Table A4 in the appendix reports bias for all estimators − expressed as a fraction of the true coefficient and multiplied by 100.[9] The numbers in brackets denote standard deviations. 'Pooled FE' denotes the fixed effects estimator applied on all clusters, and similarly for the 'pooled OLS' estimator. FE_ω denotes the fixed effects estimator applied on the true ωth cluster. As we can see, the bias of both pooled FE and OLS exceeds 20% in most cases − certainly a large number. Its negative sign is due to the fact that the clusters with small coefficients exhibit bigger leverage because the corresponding regressors have relatively large variance. The bias of the pooled OLS estimator is larger compared to pooled fixed effects because in the present simulation design the mean value of x is larger for clusters with smaller coefficients. As a result, the differences-in-variance effect

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

[9] For the pooled estimators, the true coefficient is taken to be the average value of the cluster-specific slope coefficients. The simulation set up is exactly the same as before, with the only difference being that the number of replications is $10,000$. For $\Omega_0 = 2$, $\mu_{x_1} = 1$ and $\mu_{x_2} = 4$. For $\Omega_0 = 3$, $\mu_{x_1} = 1$, $\mu_{x_2} = 4$ and $\mu_{x_3} = 5$. These differences in the mean value of the regressors influence the performance of only the pooled OLS estimator, as explained in Section 3.

on the bias is compounded by the differences-in-mean effect.[10]

6 Empirical Application

We apply the clustering approach to estimate a partially heterogeneous cost function using a panel data set of commercial banks operating in the United States. Estimation of 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 and conclude that while there is a 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 that the average cost curve has flat segments over some points, or even that the banks, depending on their size, have different average cost curves. 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.

[10] We do not report the results obtained for $K = 5$ because these are very similar to $K = 1$.

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.

Given that this view appears to have been widely adopted in the banking literature, we estimate a cost function model that allows for partial heterogeneity using the clustering criterion proposed above. A similar approach conceptually has been followed indirectly by Karapakis et al (1994), who distinguish between small and large banks and partition the population into two equally-sized subsamples. 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 1,350 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).[11] The dimensions of the data set are typical in financial and microeconomic applications. The large size of N implies that the practice of restricting the slope coefficients to be the same 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.

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 described below[12]:

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

[12] All variables are expressed in logs.

- c denotes the sum of the cost related to the three input prices that appear below;
- y_1 denotes the sum of industrial and commercial loans, real estate loans and other loans and leases;
- y_2 denotes loans to individuals;
- y_3 denotes all other assets;
- p_l denotes the price of labour, measured as total expenses on salaries and employee benefits, divided by the total number of employees;
- p_k denotes the price of capital, measured as expenses on premises and equipment, divided by the dollar value of premises and equipment; and
- p_f denotes 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:

$$\begin{aligned} c_{\omega it} &= \sum_{k=1}^{3} \beta_{k\omega} \cdot y_{k,\omega it} + \gamma_{1\omega} \cdot p_{f,\omega it} + \gamma_{2\omega} \cdot p_{l,\omega it} + \gamma_{3\omega} \cdot p_{k,\omega it} + u_{\omega it}, \\ u_{\omega it} &= \alpha_{\omega i} + \varepsilon_{\omega it} \end{aligned} \tag{34}$$

6.4 Methodology and Results

We cluster the sample of banks into up to seven clusters using the algorithm analysed 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. One empirical issue that we came across is that of scaling. In particular, we must note that the result in Section 4 is asymptotic, i.e. it applies as N gets large regardless of the level of variability in the data. However, in practice for a given sample size the level of spread in the data has to be taken into account, either through scaling the data or incorporating the scale into the θ_N term. This is because the penalty term in the clustering criterion given by (30) is a fixed number for a given N and Ω while the RSS value is scale dependent. Therefore, the clustering results may not be meaningful unless careful consideration is paid on the scale of the data.[13] This is also exemplified in our model through the

[13] The log-likelihood is also scale-dependent of course.

vast difference in the RSS value obtained by estimating the model in levels and in logs. In the former case, the RSS value is of the order of 10^{15}, which is just too large in comparison with the penalty to obtain any sensible results. To overcome this difficulty, we divide the residuals by the within standard deviation of the dependent variable, given by $sd_{within,i} = \sqrt{\frac{\sum_t (y_{it}-\bar{y}_i)^2}{T-1}}$, and then use the standardised residuals to compute RSS.

Table 2 shows the results for choosing the number of clusters using the criterion proposed in this paper, MIC, and also the Akaike, Schwarz, and Shao and Wu (2005) information criteria described in the previous section. According to our criterion, the preferred number of clusters equals three. In contrast, both Akaike and Schwarz select the maximum possible number of clusters. This is because the penalty imposed by these two criteria for over-fitting is too small in this case. Finally, SW indicates the presence of four or five clusters. Given the superior performance of MIC relative to the other criteria in the simulations, we proceed by imposing three clusters hereafter.

Table 2. Results for selecting the number of clusters.

Ω	1	2	3	4	5	6	7
MIC	642.3	393.7	**371.6**	376.7	398.1	404.6	426.9
Ak	613.3	335.8	284.8	260.9	253.4	230.8	**224.2**
SIC	610.2	329.6	275.5	248.5	237.9	212.3	**202.6**
SW	621.7	352.6	309.9	**294.4**	295.2	281.1	282.8

Table 3 reports the estimation results for the model in (34). We adopt a notation similar to the simulation study, so that $\overline{FE}$ is the average FE estimate of all clusters. We can see that there are some large differences in the value of the coefficients across clusters, which are statistically significant. For example, the cost function for the third cluster (large banks) appears to be much less responsive to changes in commercial loans and all other assets compared to the first and second clusters (small and medium banks, respectively). Furthermore, large banks appear to be more sensitive to changes in the price of labour compared to banks of smaller size. This indicates that banks of different size have considerably different cost drivers and therefore pooling the data and imposing homogeneity in the slope parameters across all individuals may yield misleading results.

Table 3. Estimation Results[a,b]

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\gamma}_1$	$\widehat{\gamma}_2$	$\widehat{\gamma}_3$
pooled FE	.179	.155	.431	.385	.282	.014
	(.004)	(.004)	(.004)	(.006)	(.003)	(.006)
pooled OLS	.396	.107	.426	.328	.159	.021
	(.004)	(.003)	(.004)	(.006)	(.003)	(.004)
FE_1	.153	.255	.522	.357	.197	-.007
	(.006)	(.006)	(.005)	(.007)	(.004)	$(.006)^*$
FE_2	.599	.049	.269	.586	.106	.079
	(.006)	(.003)	(.004)	(.006)	(.003)	(.006)
FE_3	.046	.087	.139	.409	.682	.025
	(.004)	(.005)	(.004)	(.005)	(.004)	(.004)
$\overline{FE}$	.266	.130	.310	.451	.329	.032
	(.005)	(.004)	(.004)	(.005)	(.003)	(.005)

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

7 Concluding Remarks

Complete homogeneity versus full parameter heterogeneity is an issue that has intrigued research in the analysis of panel data over the last few decades at least. In many circumstances the issue has remained 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 bridges the gap between complete homogeneity and full heterogeneity and proposes a modelling framework that imposes only partially heterogeneous restrictions in the parameters, based on the concept of 'partitional clustering'. To this end, the population of cross-sections is grouped into a fixed number of clusters using an initial partition, and then each individual is re-allocated until there is no other improvement over minimising the residual sum of squares of the estimated model. The number of clusters is determined from the model using an information criterion that is shown to be strongly consistent for fixed T. The partitional clustering algorithm is available on the web.

References

[1] Bai, Z.D., C.R. Rao and Y. Wu (1999) "Model Selection with Data-Oriented Penalty", Journal of Statistical Planning and Inference, 77, 103-117.

[2] Baltagi, B.H. and J.M. Griffin (1997) "Pooled estimators vs. their heterogeneous counterparts in the context of dynamic demand for gasoline", Journal of Econometrics, 77, 303-327.

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

[4] Baltagi, B.H., Bresson G. and A. Pirotte (2002) "Comparison of forecast performance for homogeneous, heterogeneous and shrinkage estimators. Some empirical evidence from US electricity and natural-gas consumption", *Economics Letters*, 76, 375-382.

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

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

[7] Durlauf, S and P. Johnson (1995) "Multiple regimes and cross-country growth behaviour", *Journal of Applied Econometrics,* 10, 365–384.

[8] Everitt, B. (1993) "Cluster analysis", 3rd edn., London: Eward Arnold.

[9] Hancock, D. (1986) "A Model of Financial Firm with Imperfect Asset and Deposit Elasticities", Journal of Banking and Finance, 10, 37-54.

[10] Kapetanios G. (2006) "Cluster Analysis of Panel Datasets Using Non-Standard Optimisation of Information Criteria", *Journal of Economic Dynamics and Control*, Vol. 30(8), 1389-1408.

[11] Kaparakis, E. Miller, S. and A. Noulas (1994) "Short-Run Cost Inefficiency of Commercial Banks: A Flexible Frontier Approach", *Journal of Money, Credit and Banking,* Vol. 26 No. 4, 875-893.

[12] Kaufman, L. and P.J. Rousseeuw (1990). Finding groups in data: An introduction to cluster analysis NY: John Wiley & Sons.

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

[14] McAllister, P.H. and D. McManus (1993) "Resolving the Scale Efficiency Puzzle in Banking", *Journal of Banking and Finance*, 17, 389-405.

[15] Pesaran, H. Shin Y. and R. Smith (1999) "Pooled Mean Group Estimation of Dynamic Heterogeneous Panels", *Journal of the American Statistical Association*, 94, 621-634.

[16] Rota, G.-C., (1964) "The Number of Partitions of a Set", *American Mathematical Monthly*, 71(5), 498-504.

[17] Vahid, F. (1999) "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.

[18] Zahn, C.T. (1971) "Graph-theoretical Methods For Detecting and Describing Gestalt Clusters", IEEE Trans. on Computers, vol. C-20, 68–86.

Appendices

A Proof of Equation 11

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}\mathbf{X}'_{\omega i}\mathbf{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}\mathbf{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}\mathbf{X}'_{\omega i}\mathbf{e}_T\alpha_{\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}\mathbf{X}'_{\omega i}\mathbf{X}_{\omega i}\boldsymbol{\eta}_\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(\mathbf{X}'_{\omega i}\mathbf{X}_{\omega i}\pm T\,\overline{\mathbf{X}}'_{\omega i}\overline{\mathbf{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(\mathbf{X}'_{\omega i}\mathbf{X}_{\omega i}\pm T\,\overline{\mathbf{X}}'_{\omega i}\overline{\mathbf{X}}_{\omega i}\right)\right)\boldsymbol{\eta}_\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}\mathbf{X}'_{\omega i}\mathbf{Q}_T\mathbf{X}_{\omega i}+\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}T\,\overline{\mathbf{X}}'_{\omega i}\overline{\mathbf{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}\mathbf{X}'_{\omega i}\mathbf{Q}_T\mathbf{X}_{\omega i}+T\frac{1}{N_\omega}\sum_{i=1}^{N_\omega}\overline{\mathbf{X}}'_{\omega i}\overline{\mathbf{X}}_{\omega i}\right)\boldsymbol{\eta}_\omega\right] \\
= \;& \left[\sum_{\omega=1}^{\Omega}\left(\widetilde{\mathbf{Q}}_{XX,\omega}+T\widetilde{\mathbf{M}}_{XX,\omega}\right)c_\omega\right]^{-1}\left[\sum_{\omega=1}^{\Omega}\left(\widetilde{\mathbf{Q}}_{XX,\omega}+T\widetilde{\mathbf{M}}_{XX,\omega}\right)c_\omega\boldsymbol{\eta}_\omega\right]. \qquad (35)
\end{aligned}
$$

B Proof of Theorem 1

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

We need to show that $F_N\left(\Pi_\Omega\right)-F_N\left(\Pi_0\right)>0$ a.s. for large N. Note that $\left(\psi\left(\Omega\right)-\psi\left(\Omega_0\right)\right)>0$ and, under the conditions of the theorem, θ_N grows faster than $\log\log N$.

Thus, arguing as in Shao and Wu (2005), we have

$$
\begin{aligned}
& F_N(\Pi_\Omega) - F_N(\Pi_0) \\
= \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{\mathbf{X}}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right\} + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
\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{\mathbf{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{\mathbf{X}}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right\} \\
& + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\mathbf{Y}}'_{C_{\Omega\omega}\cap C_{0,j}}\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{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(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\right)\widetilde{\mathbf{Y}}_{C_{0,j}}\right\} \\
& + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \; & T^{-1}\left\{\sum_{\omega=1}^{\Omega}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{\Omega\omega}\cap C_{0,j}}\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{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(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}\right\} \\
& + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \; & T^{-1}\left\{\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}'_{C_{0,j}}\mathbf{P}_{\widetilde{\mathbf{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}}\mathbf{P}_{\widetilde{\mathbf{X}}_{C_{\Omega\omega}\cap C_{0,j}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}\right\} + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \; & T^{-1}\left\{\sum_{j=1}^{\Omega_0}\boldsymbol{\varepsilon}'_{C_{0,j}}\mathbf{P}_{\widetilde{\mathbf{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}}\mathbf{P}_{\widetilde{\mathbf{X}}_{C_{\Omega\omega}\cap C_{0,j}}}\boldsymbol{\varepsilon}_{C_{\Omega\omega}\cap C_{0,j}}\right\} + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N,
\end{aligned}
\tag{36}
$$

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

$$
\mathbf{Q}'_{C_{0,j}}\mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\mathbf{Q}_{C_{0,j}} = \mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}. \tag{37}
$$

Under the conditions of the theorem, using (23), we have

$$
\boldsymbol{\varepsilon}'_{C_{\Omega\omega}\cap C_{0,j}}\mathbf{P}_{\widetilde{\mathbf{X}}_{C_{\Omega\omega}\cap C_{0,j}}}\boldsymbol{\varepsilon}_{C_{\Omega\omega}\cap C_{0,j}} = O\left(\log\log N_{\omega|j}\right) = O\left(\log\log N\right) \text{ a.s.}, \tag{38}
$$

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: $\Omega < \Omega_0$.

The following lemma is necessary for our proof.

Lemma 3 *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*

$$
|C_{\Omega\omega}\cap C_{0,\omega_1}| > c_0 N \text{ and } |C_{\Omega\omega}\cap C_{0,\omega_2}| > c_0 N \text{ for any } \omega \text{ and } N \text{ large enough}, \tag{39}
$$

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

$$
c_0 N < |C_{\Omega 1}\cap C_{0,1}| < N \text{ and } c_0 N < |C_{\Omega 1}\cap C_{0,2}| < N, \tag{40}
$$

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

$$\begin{aligned}
& F_N(\Pi_\Omega) - F_N(\Pi_0) \\
= \ & T^{-1}\left(\sum_{\omega=1}^{\Omega}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{\mathbf{X}}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right) + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \ & T^{-1}\left(\left\|\widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,1}} - \widetilde{\mathbf{X}}_{C_{\Omega 1}\cap C_{0,1}}\widehat{\boldsymbol{\beta}}_{\Omega 1}\right\|^2 + \left\|\widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,2}} - \widetilde{\mathbf{X}}_{C_{\Omega 1}\cap C_{0,2}}\widehat{\boldsymbol{\beta}}_{\Omega 1}\right\|^2\right) + \\
& T^{-1}\left(\sum_{\mathcal{L}_{\overline{12}}}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_{0,j}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}\cap C_{0,j}}\widehat{\boldsymbol{\beta}}_{\Omega\omega}\right\|^2 - \sum_{j=1}^{\Omega_0}\left\|\widetilde{\mathbf{Y}}_{C_{0,j}} - \widetilde{\mathbf{X}}_{C_{0,j}}\widehat{\boldsymbol{\beta}}_{0j}\right\|^2\right) + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N.
\end{aligned} \tag{41}$$

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

$$\widetilde{\mathbf{Y}}_{012} = \begin{pmatrix} \widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,1}} \\ \widetilde{\mathbf{Y}}_{C_{\Omega 1}\cap C_{0,2}} \end{pmatrix},\ \widetilde{\mathbf{X}}_{012} = \begin{pmatrix} \widetilde{\mathbf{X}}_{11} \\ \widetilde{\mathbf{X}}_{12} \end{pmatrix},\ \widetilde{\boldsymbol{\varepsilon}}_{012} = \begin{pmatrix} \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega 1}\cap C_{0,1}} \\ \widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega 1}\cap C_{0,2}} \end{pmatrix}. \tag{42}$$

Hence

$$\begin{aligned}
& F_N(\Pi_\Omega) - F_N(\Pi_0) \\
\geq \ & T^{-1}\left(\left\|\widetilde{\mathbf{Y}}_{012} - \widetilde{\mathbf{X}}_{012}\widehat{\boldsymbol{\beta}}_{012}\right\|^2 + \sum_{\mathcal{L}_{\overline{12}}}\left\|\widetilde{\mathbf{Y}}_{C_{\Omega\omega}\cap C_{0,j}} - \widetilde{\mathbf{X}}_{C_{\Omega\omega}\cap C_{0,j}}\widehat{\boldsymbol{\beta}}_{\omega|j}\right\|^2\right) \\
& -T^{-1}\sum_{j=1}^{\Omega_0}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}}'\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N,
\end{aligned} \tag{43}$$

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

$$\begin{aligned}
& F_N(\Pi_\Omega) - F_N(\Pi_0) \\
\geq \ & T^{-1}\left(\widetilde{\mathbf{Y}}_{012}'\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{012}}\right)\widetilde{\mathbf{Y}}_{012} + \sum_{\mathcal{L}_{\overline{12}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}'\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{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(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \ & T^{-1}(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02})'\widetilde{\mathbf{X}}_{11}'\left[\mathbf{I} - \widetilde{\mathbf{X}}_{11}\left(\widetilde{\mathbf{X}}_{11}'\widetilde{\mathbf{X}}_{11} + \widetilde{\mathbf{X}}_{12}'\widetilde{\mathbf{X}}_{12}\right)^{-1}\widetilde{\mathbf{X}}_{11}'\right]\widetilde{\mathbf{X}}_{11}(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}) \\
& +T^{-1}2(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02})'\widetilde{\mathbf{X}}_{11a}'\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{012}}\right)\widetilde{\boldsymbol{\varepsilon}}_{012} + T^{-1}\widetilde{\boldsymbol{\varepsilon}}_{012}'\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{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(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{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(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\right)\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} \\
& + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N \\
= \ & T^{-1}(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02})'\left[\left(\widetilde{\mathbf{X}}_{11}'\widetilde{\mathbf{X}}_{11}\right)^{-1} + \left(\widetilde{\mathbf{X}}_{12}'\widetilde{\mathbf{X}}_{12}\right)^{-1}\right]^{-1}(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02}) + \\
& T^{-1}2(\boldsymbol{\beta}_{01} - \boldsymbol{\beta}_{02})'\widetilde{\mathbf{X}}_{11a}'\left(\mathbf{I} - \mathbf{P}_{\widetilde{\mathbf{X}}_{012}}\right)\widetilde{\boldsymbol{\varepsilon}}_{012} - T^{-1}\widetilde{\boldsymbol{\varepsilon}}_{012}'\mathbf{P}_{\widetilde{\mathbf{X}}_{012}}\widetilde{\boldsymbol{\varepsilon}}_{012} \\
& -T^{-1}\sum_{\mathcal{L}_{\overline{12}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{\Omega\omega}\cap C_{0,j}}'\mathbf{P}_{\widetilde{\mathbf{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}}'\mathbf{P}_{\widetilde{\mathbf{X}}_{C_{0,j}}}\widetilde{\boldsymbol{\varepsilon}}_{C_{0,j}} \\
& + (\psi(\Omega) - \psi(\Omega_0))\,\theta_N,
\end{aligned} \tag{44}$$

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

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

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

Using (21), (23) and the Cauchy-Schwartz inequality we see that the other terms that do not involve θ_N in the above lower bound are of smaller order in N. In this case, $(\psi(\Omega) - \psi(\Omega_0)) < 0$. Thus, for N large enough, $F_N(\Pi_\Omega) - F_N(\Pi_0) > 0$ a.s. since, by assumption, $\lim_{N\to\infty} N^{-1}\theta_N = 0$.

Table A1. Simulation results for selecting the number of clusters when $\Omega_0 = 1$.

	$K=1$				$K=5$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
MIC								
$\Omega=1$	**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
Ak								
$\Omega=1$	**.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
SIC								
$\Omega=1$	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
$\Omega=2$	.760	.802	.760	.802	.760	.802	.760	.802
$\Omega=3$	.240	.198	.240	.198	.240	.198	.240	.298
SW								
$\Omega=1$	**.000**	**.002**	**.000**	**.002**	**.000**	**.002**	**.000**	**.002**
$\Omega=2$	.982	.964	.982	.964	.994	.962	.994	.962
$\Omega=3$	.018	.034	.018	.034	.006	.036	.006	.036

Table A2. Simulation results for selecting the number of clusters when $\Omega_0 = 2$.

	$K=1$				$K=5$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
MIC								
$\Omega=1$	.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
Ak								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
$\Omega=3$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=4$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SIC								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	**.046**	**.072**	**.068**	**.086**	**.034**	**.064**	**.072**	**.086**
$\Omega=3$	.944	.926	.926	.908	.952	.928	.922	.914
$\Omega=4$	.010	.002	.06	.000	.014	.008	.026	.000
SW								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	**.264**	**.312**	**.302**	**.336**	**.250**	**.962**	**.412**	**.444**
$\Omega=3$	.724	.684	.686	.664	.742	.038	.586	.556
$\Omega=4$	.012	.004	.012	.000	.08	.004	.002	.000

Table A3. Simulation results for selecting the number of clusters when $\Omega_0 = 3$.

	$K=1$				$K=5$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
MIC								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.072	.044	.062	.038	.062	.034	.058	.028
$\Omega=3$	**.928**	**.956**	**.938**	**.962**	**.938**	**.966**	**.942**	**.972**
$\Omega=4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=5$	.000	.000	.000	.000	.000	.000	.000	.000
Ak								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=3$	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**	**.000**
$\Omega=4$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=5$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SIC								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=3$	**.764**	**.826**	**.836**	**.850**	**.742**	**.776**	**.842**	**.854**
$\Omega=4$	.156	.164	.158	.150	.158	.218	.154	.142
$\Omega=5$	.080	.040	.006	.000	.100	.006	.004	.004
SW								
$\Omega=1$	.000	.000	.000	.000	.000	.000	.000	.000
$\Omega=2$	.000	.000	.012	.006	.000	.000	.012	.004
$\Omega=3$	**.782**	**.802**	**.874**	**.876**	**.770**	**.776**	**.902**	**.906**
$\Omega=4$	.218	.198	.126	.124	.226	.224	.098	.094
$\Omega=5$	.000	.000	.000	.000	.004	.000	.000	.000

Table A4. Finite sample properties of estimators.

	$K=1$, $\Omega_0=2$, $\overline{\beta}=0.75$				$K=1$, $\Omega_0=3$, $\overline{\beta}=0.125$			
	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$	$N=100$	$N=400$
	$\zeta=4$		$\zeta=8$		$\zeta=4$		$\zeta=8$	
pooled FE	−18.95	−20.12	−18.95	−20.11	−38.04	−37.60	−44.25	−44.25
	(.011)	(.005)	(.008)	(.004)	(.003)	(.002)	(.002)	(.001)
pooled OLS	−24.14	−24.61	−22.61	−23.16	−53.64	−60.46	−57.71	−51.36
	(.019)	(.009)	(.013)	(.006)	(.005)	(.003)	(.003)	(.002)
FE_1	.060	−.025	.040	−.017	.368	.058	−.062	−.044
	(.024)	(.012)	(.017)	(.009)	(.015)	(.007)	(.011)	(.005)
FE_2	−.010	−.084	−.006	−.058	−.192	.096	0.024	0.016
	(.012)	(.006)	(.009)	(.004)	(.004)	(.002)	(.003)	(.001)
FE_3					.428	.020	.004	.004
					(.008)	(.004)	(.006)	(.003)