pp. 1–48.

Identification and Estimation of Dynamic Heterogeneous Unbalanced Panel Data Models with Clustering

Updated December 29, 2023

Monika Avila Márquez[†]

[†] *School of Economics,*
University of Bristol, 12A Priory Rd, Bristol BS8 1TU, United Kingdom.
E-mail: monika_avilam@hotmail.com

Summary This paper investigates the identification and estimation of dynamic heterogeneous linear models for unbalanced panel data at the cluster level when the clustering structure is known, and the number of time periods is short (greater than or equal to 3). For this purpose, we use a linear panel data model with additive cluster fixed effects and a mixed coefficient structure composed of cluster-specific fixed effects and random cluster-individual-time specific effects. We propose a Mean Cluster-FGLS estimator and a Mean Cluster-OLS estimator to estimate the mean coefficients. In order to make the GLS estimation of the cluster-specific parameters feasible, we introduce a ridge estimator of the variance-covariance matrix of the model. The Mean Cluster estimators are consistent: i) under stratified sampling when the number of clusters is fixed, the proportion of observed clusters is equal to 1, and the number of individuals per cluster grows to infinity, or ii) under cluster sampling when the square root of the number of clusters grows at a slower rate than the growth rate of the number of individuals per cluster. In addition, we present two extensions of the baseline model. In the first one, we allow for cluster-individual specific fixed effects instead of cluster additive fixed effects. In this setting, we propose a Hierarchical Bayesian estimator that considers the problem of unknown initial conditions. In the second extension, we allow for cross-sectional dependence by including common factors. We propose the Mean Cluster estimator using the time-demeaned variables to estimate this model.

1. INTRODUCTION

Heterogeneous linear dynamic panel data models with short time dimension (T) suffer from two well-known problems: the incidental parameter bias (Nickell (1981)) and the unknown initial conditions dependency (Hsiao (2020), Wooldridge (2005b)).

The first-difference GMM estimator (Arellano and Bond (1991)) is inconsistent when the persistence parameter is heterogeneous across individuals, t. The reason is that ignoring the individual heterogeneity in the persistence parameter is equivalent to model misspecification. If we assume that the individual persistence parameter is random, we end up with an endogeneity issue without available instrumental variables. In a heterogeneous dynamic panel data model with a long time dimension, one can use the Mean Grop estimator. But when the time dimension is as short as 3, the Mean Group estimator is unfeasible because the risk of the individual least squares estimator is of order K/T, with K representing the total number of covariates. Another limitation of the mean group estimation is that the small number of time observations prevents including a big number of covariates.

While the identification and estimation of dynamic linear panel data models with unobserved multiplicative individual heterogeneity and time dimension as short as 3 seems hopeless, one can still find a workaround for the problem if individuals present similar behavior within known clusters. A known clustering structure is possible in sampling frameworks where the population is clearly clustered. For instance, one could think about households within counties, employees within firms, firms within industries, etc.

This motivates the proposal of an alternative estimation methodology for dynamic linear heterogeneous panel data models that exploits the clustering structure in the data. For this purpose, we assume that the persistence coefficient and the additive individual heterogeneity are homogeneous within cluster, and that the multiplicative individual unobserved heterogeneity is partitioned into two components, multiplicative individual heterogeneity correlated with the regressors that is pooled within clusters and multiplicative individual heterogeneity that is uncorrelated with the regressors within clusters. Under these key assumptions, it is possible to obtain consistent estimates and overcome the incidental parameter bias as well as the initial conditions problem.

The heterogeneity is modeled with a mixed coefficient structure composed of fixed cluster-specific effects and random cluster-individual-time-specific effects. Therefore, the model considered in this paper presents additive and multiplicative cluster fixed effects instead of individual-specific fixed effects.

The key assumption of a mixed coefficient structure is related, but not equal, to the assumption presented by Krishnakumar et al. (2017) for a static three-level linear panel data model. The latter assumption states that the coefficient vector is equal to the sum of a mean coefficient vector plus fixed specific effects and random specific effects. In contrast, the former assumption states that the coefficient vector equals the sum of varying coefficients at cluster level plus cluster-individual-time random components. In addition, the assumption of a mixed coefficient structure is related to the assumption described by Hsiao (2014) for two-level panel data that states that coefficients are composed of a systematic component driven by observed regressors and a random component.

The advantage of including cluster fixed effects instead of individual fixed effects is that the number of clusters specific fixed effects is lower. The dimensionality reduction of the fixed effects allows consistent estimation because the problem of incidental parameter bias disappears when the number of individuals in the cluster n_g grows. Another advantage is that the initial condition dependency is controlled. In contrast, a disadvantage of including cluster fixed effects instead of individual effects is that the model is misspecified if individuals do not pool within clusters. We address this problem by extending the model and allowing additive cluster-individual fixed effects. Another problem surges if the assumed clustering structure is not correct.

More specifically, we investigate the identification and estimation of dynamic heterogeneous linear models for clustered panel data that is unbalanced due to randomly missing data and with short time dimension. For this purpose, we use a three-dimensional panel data framework and consider the following baseline model for individual i belonging to cluster g:

$$y_{git} = \rho_g y_{git-1} + x'_{git}\beta_{git} + \alpha_{1,g} + \varepsilon_{git}, \qquad t = 1, 2, ..., T_{i_g}. \tag{1.1}$$

Index i refers to individual i belonging to cluster g, index t refers to the time observation t of individual i belonging to cluster g. [1] The number of observed groups in the panel

[1] We could have used the alternative notation i_g that represents individual i belonging to cluster g and

is equal to m, the number of individuals per group equals to N_g, and the number of observations per individual i in group g is equal to T_{i_g}. Each group g has a total number of observations equal to $n_g = \sum_{i_g} T_{i_g}$.
The parameters of interest of model 1.1 are the cluster specific persistence parameter ρ_g and the cluster specific mean coefficients ($\beta_g = E[\beta_{git}|\mathcal{F}_g]$ with $\mathcal{F}_g$ representing cluster g sub-sigma-field), as well as their overall averages.

Additionally, we allow for residual random multiplicative cluster-individual-time specific heterogeneity in the coefficient vector that aims to capture possible random deviations of individuals from their cluster.

To estimate the baseline model 1.1, we propose two Mean Cluster (MC) estimators and the cluster-specific estimators. The Mean Cluster estimators are the mean of the FGLS or OLS parameter estimations of each cluster g. In order to make GLS feasible, we propose a ridge estimation of the variance-covariance components along with a modification suitable for a big sample size. These estimators are consistent: i) under stratified sampling when the number of clusters is fixed, the proportion of observed clusters is equal to 1 and the number of individuals per cluster grows to infinity or ii) under cluster sampling when the square root of the number of clusters grows at a slower rate than the growth rate of the number individuals per cluster.

The main advantages of the Mean Cluster estimators are: i) the estimation of dynamic heterogeneous panel data models with only three-time observations is possible, ii) the cluster-specific persistence parameters are identified, iii) the number of covariates included in the model is not restricted by the size of the time dimension, and iv) the computational burden is lower since one partitions the data in clusters. The latter happens because the estimation technique performs a first step local optimization and global optimization when averaging in the second step. The main disadvantages of the Mean Cluster FGLS or OLS estimators are i) not robust to violation of cluster assumption, and ii) inconsistent when the proportion of observed clusters is lower than 1 under stratified sampling.

In order to test the assumption of clustered individual heterogeneity, we propose two specification tests that are extensions of the Hausman test (Hausman and Taylor, 1981). First, testing the null hypothesis of cluster additive and multiplicative heterogeneity versus cluster-individual additive and multiplicative heterogeneity is not feasible when the time dimension is as short as 3. But, testing the null hypothesis of complete homogeneity versus cluster additive and multiplicative heterogeneity is possible. In this case, we propose to compare the Mean Cluster estimator with the simple Pooled OLS estimator. In addition, testing the null hypothesis of cluster additive and multiplicative effects versus cluster-individual additive heterogeneity and cluster multiplicative heterogeneity is also viable. In this case, we propose to compare the Mean Cluster estimator against the Mean Cluster first-difference GMM estimator or the Mean Cluster estimator using a Mundlak approach. The study of the statistical properties of these tests is left for further research.

It is clear that the failure of the assumption of clustered heterogeneity causes inconsistency of the estimators. As a possible solution, we extend the baseline model 1.1 to allow for the presence of cluster-individual specific additive effects. In this setting, we are back to the problems of incidental parameters and initial conditions dependency. In order to deal with the incidental parameters, we use the Mundlack approach and we propose

t_{i_g} for time observation t of individual i belonging to group g as explained in Section 2. But we use three indexes to simplify the notation.

a Bayesian hierarchical estimator with a prior for the initial conditions. The Bayesian estimator requires the correct specification of the prior of the initial conditions. Thus, the assumption of initial conditions generated from the stationary distribution is critical for consistency of the proposed Bayesian estimator. While we present an alternative prior allowing for initial conditions that are not generated from the stationary distribution, it is not straightforward to decide which is the correct assumption for the initial conditions. As an alternative, we use the Chamberlain or the Mundlak approach conditioning on the initial values (Wooldridge (2005b)) and we propose to estimate the the cluster-specific parameters using a factor analytical method following Bai (2013). Another issue is the potential cross-sectional dependence within clusters. In order to deal with this problem, we extend the baseline model 1.1 to a setting that includes common factors and we propose Mean Cluster estimation using the time-demeaned variables (Sarafidis and Robertson (2009)).

The literature for dynamic heterogeneous linear panel data models focuses on two-level panel data models or models that ignore clustering. Pesaran et al. (1999) proposes a Mean Group estimator that averages the OLS estimators for each individual in the panel. This estimator is consistent when the time dimension grows to infinity and needs debiasing when the time dimension is short. Hsiao et al. (1998) presents a hierarchical Bayes estimator for small panels that assumes that the initial conditions are fixed. The literature for clustering in panel data concentrates on panels with a long-time dimension. Bester and Hansen (2016) propose a grouped estimator for fixed effects non-linear models based on observable characteristics. Bonhomme and Manresa (2015) propose a grouped fixed effects estimator that converges to a pseudo true value that it is not necessarily equal to the true value when the time dimension is as short as 3 (Sarafidis and Wansbeek (2021)).

This paper contributes to the literature in five ways: i) it introduces an assumption of a mixed coefficient structure for three-level panel data that states that the coefficients are composed of fixed coefficients varying at the cluster level and cluster-individual specific random effects and that is appropriate for a setting under stratified sampling [2], ii) it proposes a Mean Cluster estimator, iii) it provides the conditions for consistency and asymptotic normality of the Mean Cluster estimators under stratified and cluster sampling, iv) it provides an estimation method for the variance-covariance of the model by extending the methodology presented by Krishnakumar et al. (2017) to a dynamic setting, v) it proposes a hierarchical Bayesian estimator that takes into account the initial conditions, vi) it shows that the first-difference GMM estimator is inconsistent in the presence of cluster heterogeneity in the persistence parameter. In addition, the paper presents a discussion about violations of the assumed variance-covariance matrix of the model, it proposes to condition on the initial values to relax the assumption that the initial conditions are generated from the stationary distribution (Wooldridge (2005b), Hsiao et al. (2002)), it proposes specification tests, it shows that when the time dimension is long the Mean Group estimator is equivalent to the Mean Cluster estimator under stratified sampling, and it discusses a setting with long time dimension.

The rest of the paper is organized as follows: Section 2 explains the structure of the

[2] This assumption is not equal to the one proposed by Hsiao et al. (1989). The authors proposed mixed fixed and random coefficients, which means that some regressors present fixed coefficients and other random coefficients. In contrast, I assume that the coefficients of the regressors are the sum of cluster fixed specific effects and random effects.

data, Section 3 presents the model with its assumptions and its relationship with two-level panel data, Section 5 states the identification strategy of the parameters of interest, Section 4 presents the estimation strategy, Section 6 exposes the statistical properties of the methods proposed, Section 7 discusses consequences of misspecification of the variance-covariance matrix of the disturbance term of the model, 8 compares the Mean Cluster estimator with other available estimators, Section 9 presents the necessary assumptions for a setting under cluster sampling, Section 10 explains possible limitations of model 1.1, Section 11 presents specification tests, Section 12 relaxes the assumption of additive cluster effects to cluster-individual additive specific effects and presents Bayes estimation and estimation conditioning on the initial values, Section 13 presents an extension of the model with cross sectional dependence, Section 14 discusses the behavior of the Mean Cluster estimator in a setting with long time dimension, Section 15 explains the challenges of unknown clustering when the time dimension is as short as 3, Section 17 describes the Monte Carlo experiments and the results, Section 18 gives the conclusions.

Notation: $||\cdot||^2$ is the Euclidean norm. $||\cdot||_F$ is the Frobenius norm. Scalar random variables are collected in column vectors; for instance y_{git} can be collected in the vector $y \in \mathbb{R}^M$ ($y = [y_{111} \quad ... \quad y_{mN_mT_{i_m}}]'$). Matrices are denoted by uppercase letters; for instance the matrix $X \in \mathbb{R}^{n\times K}$ that collects the transpose of the column vector $x_{git} \in \mathbb{R}^{K\times 1}$ containing K regressors corresponding to individual i belonging to cluster g at period t. I_A represents the identity matrix with dimension $A \times A$ where A is a positive integer. $\mathbf{0}$ represents a vector of zeros with dimensions $K \times 1$.

2. DATA STRUCTURE

The data $\{y_{it}, x_{it}\}_{i=1}^{N}$ is obtained from stratified sampling, and it can be partitioned in m nonoverlapping subsets $\{y_{i_gt}, x_{i_gt}\}_{i_g=1}^{N_g}$. The population is stratified in m nonoverlapping independent known clusters, this means that the number of observed clusters m is equal to the total number of clusters in the population. In contrast, under cluster sampling the number of observed clusters m is not equal to the total number of clusters in the population. Individuals are independent within cluster (this is relaxed in Section 13). For each cluster g, N_g individuals are sampled over T_{i_g} periods. The total number of individuals across clusters is $N = \sum_g^m N_g$. The total number of observations per cluster g is $n_g = \sum_{i_g} T_{i_g}$. The total number of observations in the data set is $n = \sum_g^m n_g$. This data can be seen as an unbalanced three-level panel.

We define the following subscripts:

- g denotes each group and takes values $g \in \{1, 2, ..., m\}$.
- i_g denotes individual i_g in group g and takes values $i_g \in \{1, 2, ..., N_g\}$.
- t_{i_g} denotes time observation t of individual i_g in group g and takes values $t_{i_g} \in \{1, 2, ..., T_{i_g}\}$.

Remark 2.1. For simplicity, we use i and t equivalently to i_g and t_{i_g}. This does not mean that we assume that individual i is not subordinated to g.

3. THE MODEL

We consider the autoregressive distributed lag ARDL(1,0) heterogeneous panel data model for a random draw i from the population of cluster g:

$$y_{git} = \alpha_{1,g} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git}, \qquad t = 1, ..., T_{i_g}, \tag{3.2}$$

with:

$$\beta_{git} = \beta_g + \lambda_{git}. \tag{3.3}$$

where y_{git} is the observed outcome variable with support $\mathbb{Y} \subseteq \mathbb{R}$, y_{git-1} is the first lag of the outcome variable, and x_{git} is a $K \times 1$ vector of observed explanatory variables for individual i in cluster g for period t with support $\mathbb{X} \subseteq \mathbb{R}^K$ (variables with finite support are also allowed), ε_{git} is an unobserved idiosyncratic cluster-individual error term in period t.

The unobserved parameters of interest are the cluster-specific parameter (ρ_g) and the cluster-specific slope coefficients (β_g). The model also includes cluster additive specific fixed effects ($\alpha_{1,g}$) as well as multiplicative cluster-individual-time specific effects (λ_{git}). Since individuals belong to an overall population that is partitioned in known clusters, there is also interest in the overall averages of the parameters $E[\rho_g]$, $E[\beta_g]$. [3]

The total number of time observations per individual T_{i_g} is small and considered fixed in the asymptotic analysis. The number of individuals per cluster is N_g and the total number of individuals in the panel N are growing to infinity. The number of clusters is fixed under stratified sampling. This setting can be evaluated using an asymptotic sequence framework where we allow N_g to grow and the time dimension T_{i_g} is fixed (Moon et al., 2018).

As mentioned before, it is well known that the growth of the individual dimension produces an incidental parameter bias when there is individual-specific heterogeneity and the time dimension is short. A standard approach to avoid this incidental parameter problem is to assume random coefficients for each individual i in the sample or allow for additive individual fixed effects. In this paper, we handle this problem by imposing clustered heterogeneity and using a novel mixed structure in the slope coefficients.

More specifically, we assume that ρ_g is fixed and the slope coefficient vector presents a mixed structure ($\beta_{git} = \beta_g + \lambda_{git}$) composed of a cluster-specific fixed component (β_g) and a random cluster-individual-time specific effect λ_{git}. In addition, we assume a full variance-covariance matrix for the random cluster-individual-time specific effect that captures the covariance between marginal effects of the included regressors in the model. This coefficient structure allows for possible clustered endogenous heterogeneity while admitting random deviations of individual time-specific marginal effects from their cluster mean. For instance, one could think that common cultural unobserved characteristics drive the heterogeneous habit formation of individuals in a certain cluster while possible deviations are random and noncorrelated to "taste-shifters". [4]

The mixed coefficient structure can have three possible interpretations: i) the data is sampled from a density function with heterogeneous parameters, ii) the correlation of the regressors with unobserved multiplicative individual heterogeneity is equal within cluster, or iii) the regressors are freely correlated to multiplicative cluster unobserved heterogeneity while preserving noncorrelation with multiplicative cluster-individual-time specific unobserved heterogeneity (See Section 9). An example of the second interpretation is that innate ability and the marginal return to education of individuals are equally correlated to education within a city if we believe that individuals with higher ability do

[3] They can be seen as average partial effects as explained by Wooldridge (2005a).

[4] Dynan (2000) calls "taste-shifter" to preference related variables.

not only self-select into education levels but also into the city where they will have the highest return to their education.

Model 3.2 is relevant for different empirical applications because it permits accounting for correlated cluster heterogeneity as well as individual and time heterogeneity. For instance, one could be interested in studying dynamic heterogeneous demand equations, the heterogeneity of habit formation, income persistence, dynamic heterogeneous treatment effects, and others.

In the following lines, we present the assumptions of the model in more detail.

Assumption 3.1. *Cluster membership is known and fixed over time.*

The researcher knows the clusters based on observed characteristics. For instance, clustering can be done by counties, sub-regions, economic activity categories at a detailed level, among others. The membership of individual i into cluster g is denoted by the indicator variable $s_i^{(g)} \in \{0,1\}$ that takes value 1 if the individual belongs to cluster g and 0 otherwise. Thus, each individual has m indicator variables. It is crucial to notice that cluster belonging does not vary with time.
The sum of $s_i^{(g)}$ for all individuals in the panel gives the number of individuals in the cluster g ($\sum_i^N s_i^{(g)} = N_g$).

Assumption 3.2. *Number of individuals within cluster is growing.*

$$N \to \infty \Rightarrow N_g \to \infty, \quad \forall g \in \{1,2,...,m\}.$$

The number of individuals within cluster grows to infinity when the number of individuals in the panel grows to infinity. This could happen for households within sub-region or enterprises in an economic sector.

Assumption 3.3. *Non vanishing clusters.*

$$\underset{N\to\infty}{lim} \frac{N_g}{N} \to \pi_g, \quad \forall g \in \{1,2,...,m\},$$
$$\pi_g \in (0,1).$$

The proportion of cluster population to the overall population converges to a fixed number greater than 0 but less than 1 as the number of individuals within cluster and the total number of individuals in the panel grows to infinity. This assumption implies that the number of clusters is fixed.
It is possible to assume that the number of clusters grows. In this case, this assumption is replaced by vanishing clusters and it is necessary to add a restriction to its growth rate by assuming that it grows at a slower rate than the squared number of individuals in the cluster such that $\frac{\sqrt{m(n_g)}}{n_g} \to 0$ as $n_g \to \infty$ (See Section 9). This means that the number of clusters is an increasing monotonic function of the total number of observations within cluster, and its square root is $o(n_g)$. An example of this setting could be the Public Use Microdata Areas (PUMA) of the USA. Each PUMA has at least 100,000 individuals per unit, and the number of PUMAs is large. In this case, we can assume that cluster-specific effects are random, either correlated or not to the regressors, and the

estimation methodology given in Section 4 is still consistent for both cluster and mean coefficients. Nevertheless, the asymptotic framework differs from the one presented in Section 6 and it is provided in Section 9. If one desires to relax completely the requirement of growing individuals per cluster and still obtain unbiased estimators per cluster, one needs a debiased cluster estimator.

Assumption 3.4. *The proportion of observed clusters (q) concerning to the total number of clusters in the population is equal to 1.*

This assumption is in line with stratified sampling. If the proportion of observed clusters (m) with respect to the total number of clusters in the population is lower than 1 and the number of clusters in the population is small, the sample is not representative of the underlying population. As a result, the Mean Cluster estimator is unfeasible as there is insufficient information. If the proportion of observed clusters (m) with respect to the total number of clusters in the population is lower than 1 and the number of clusters in the population is large, we are in a setting where not all clusters in the population are sampled and correspond to a cluster sampling setting. In Section 9, we present the assumptions compatible with cluster sampling. An example of data obtained by means of cluster sampling is the one used by Andrabi et al. (2011).

Assumption 3.5. *Fixed cluster additive specific effects $\alpha_{1,g}$.*

Assumption 3.6. *Fixed cluster specific persistence parameter.*

$$\rho_g \in (-1, 1).$$

Assumption 3.7. *Mixed cluster-individual-time specific coefficients.*

$$\beta_{git} = \beta_g + \lambda_{git},$$

$$E[\lambda_{git}\lambda'_{g'i't'}|x_{gi1}, x_{gi2}, ..., x_{giT}] \quad = \begin{cases} \Delta_{\lambda_g} & if \quad g = g', i = i' and t = t', \\ 0 & otherwise, \end{cases} \quad .$$

The unobserved coefficient vector is composed of a fixed cluster coefficient vector (β_g), and a heteroskedastic random component (λ_{git}) conditional on covariates that captures the multiplicative heterogeneity over time for each individual of cluster g.

Assumption 3.8. *The random cluster-individual-time effects have zero mean conditional on the covariates.*

$$E[\lambda_{git}|x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}] = 0.$$

This implies that $E[\beta_{git}|x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}] = \beta_g$. As a consequence of this assumption and Assumption 3.5, $E[\lambda_{git}] = 0$.

Assumption 3.9. *Strict exogeneity of the covariates with the disturbance term.*

$$E[\varepsilon_{git}|x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}] = 0.$$

This assumption is in line with Hsiao et al. (1998) and it rules out possible feedback of y_{git} with future values of the covariates. It implies the model presents dynamic completeness without conditioning on cluster effects because cluster-specific effects are considered fixed parameters. It is also possible to assume that the cluster effects are random and correlated with the regressors. With correlated cluster effects, the strict exogeneity of the covariates must be conditional on the cluster-specific effects. The orthogonality conditions presented in section 5 hold under strict exogeneity of the covariates conditional on the cluster-specific effects (See Section 9). As a consequence of this assumption and Assumption 3.5, $E[\lambda_{git}] = 0$.

Remark 3.1. According to Wooldridge (2010), strict exogeneity rules out possible feedback of the past values of the dependent variable to the covariates. Allowing for this feedback requires relaxing this assumption to sequential exogeneity. The assumption of sequential exogeneity is weaker than strict exogeneity since it allows for feedback from y_{git} to $x_{git+1}, ..., x_{giT}$. For instance, consumption in period t can have an effect on taste shifters in periods after t. In order to allow for this possible feedback, it is necessary to modify the first stage of the estimation method proposed in section 4 by replacing OLS or GLS with GMM using instrumental variables.

Assumption 3.10. *The error term ε_{git} is homoskedastic, and uncorrelated within each cluster g but heteroskedastic across clusters conditional on regressors .*

$$E[\varepsilon_{git}^2|x_{gi1}, x_{gi2}, ..., x_{giT}] = \sigma_{\varepsilon_g}^2 < \infty.$$

$$E[\varepsilon_{git}, \varepsilon_{g'i't'}|x_{gi1}, x_{gi2}, ..., x_{giT}] = 0, \quad if \quad g \neq g', i \neq i', t \neq t'.$$

Assumption 3.11. *y_{git} are generated from the stationary distribution of the process with initialization values $y_{gi,-h_{i_g}}$ sampled h_{i_g} number of periods before the data collection in period 0.*

The initial conditions are given by:

$$y_{gi0} = \rho_g^{h_{i_g}} y_{gi,-h_{i_g}} + \alpha_{1,g}\frac{1-\rho_g^{h_{i_g}}}{1-\rho_g} + \sum_{l=0}^{h_{g_i}} \rho_g^l x'_{gi-l}\beta_{gi-l} + \sum_{l=0}^{h_{i_g}} \rho_g^l \varepsilon_{gi-l}. \tag{3.4}$$

with h_{i_g} unrestricted. It is possible to set h_{i_g} free because the initial conditions (y_{gi0}) dependence is controlled under the assumption of fixed cluster additive effects (Assumption 3.5).

Assumption 3.11 is not necessary in the presence of fixed cluster-additive effects (Assumption 3.5). The reason is that the dependence of the initial conditions y_{gi0} is controlled because the cluster-additive effects are assumed to be fixed. As a result, the Mean Cluster estimator (Section 4) is consistent without Assumption 3.11 if the Assumption 3.5 holds. In contrast, Assumption 3.11 is essential if there are cluster-individual additive effects instead of cluster additive effects. The reason is that, under Assumptions 3.11 and 3.12, the initial conditions can be projected into all past, present and future values of the regressors. This means that it is possible to estimate the model either using a Bayesian approach or conditioning on the initial value y_{gi0} (Hsiao et al. (2002)).

In addition, if the model presents cluster-individual additive fixed effects instead of cluster additive effects and h_{i_g} is small, the individual initialization values $y_{gi,-h_{i_g}}$ are essential because there exist initial conditions (y_{gi0}) dependence. In that case, there is a need to add an assumption to avoid the incidental parameter problem: $E[y_{gi,-h_{i_g}}] = b_g$. On the other hand, having $h_{i_g} \to \infty$ means that the effect of the initialization value dies (Similar to Hsiao et al. (2002)).

Assumption 3.11 can be relaxed in the presence of cluster-individual additive effects. In this case, the Bayesian estimator requires a prior for the initial conditions not generated from the stationary distribution. A simpler solution is to condition on the initial conditions as suggested by Wooldridge (2005b).

Assumption 3.12. *x_{git} are generated from:*

$$x_{git} = \mu_g + \rho_x x_{git-l} + \omega_{git}, \qquad |\rho_x| < 1.$$

x_{git} are stationary with ω_{git} i.i.d with variance σ^2_ω. This assumption is similar but not equal to the one presented by Hsiao et al. (2002). Assumption 3.12 in combination with Assumption 3.6 states that the dependent variable and the regressors are both integrated of order 0.

It is possible to relax Assumption 3.12 and allow for the presence of cluster-specific trends as follows:

$$x_{git} = \mu_g + b_g t + \rho_x x_{git-l} + \omega_{git}, \qquad |\rho_x| < 1.$$

The Mean Cluster estimator presented in section 4 is consistent with trend stationary regressors if we include a deterministic trend in model 3.2. Otherwise, it is consistent only if the data-generating process started a short time ago (small h_{i_g}). An example could be the wage of young individuals, which means one could include age or experience as regressors in the model.

When the model presents cluster additive specific effects, relaxing Assumption 3.12 causes a non-stationary dependent variable. The cluster-specific estimators remain consistent if the regressors and the dependent variable are co-integrated per cluster. In addition, the inclusion of the lag of the dependent variable in the right-hand side of the equation may produce a stationary error term (Hamilton (1994)). As a result, the Mean Cluster estimator may remain consistent and asymptotically normal but this is left for further research. If the regressors and the dependent variable are not co-integrated per cluster, it is not clear if the pooled cluster OLS estimator is consistent even if Phillips and Moon (1999) show that pooled OLS is a consistent estimator of the long-run average regression coefficient if the regressors are non-stationary and there is no co-integration. The reason is that they considered a model that does not present an intercept and the lag of the dependent variable. In order to test for co-integration, one needs to extend the test proposed by Im et al. (2003) to allow for cluster-specific parameters instead of individual-specific parameters. Concluding that there is co-integration would entail that $u_{git} = x'_{git}\lambda_{git} + \varepsilon_{git}$ is stationary, implying that λ_{git} could be considered as a random co-integrating vector. A study of a co-integration test and the properties of the Mean-

Cluster estimator when there is no co-integration is outside the scope of this paper, and both issues are left for further research.

When the model presents cluster-individual specific effects instead of cluster-specific effects, the assumption of stationary regressors is important. The reason is that the presence of cluster-individual specific effects causes the problem of initial conditions dependency. This problem can be solved by projecting the initial conditions on the past, the present, and the future values of the regressors. Moreover, the projection of the initial conditions on the regressors is only possible if the regressors are stationary (Hsiao (2020)). Thus, non-stationary regressors cause the failure of the Bayes estimator proposed in section 12.1. The reason is that it is not possible to project the initial conditions on the cluster-individual mean of the regressors Hsiao (2020). A solution is conditioning on the initial conditions as proposed by Wooldridge (2005b) because this does not require Assumption 3.12. In this case, the Mean Cluster estimator is consistent in the presence of non-stationary regressors with or without co-integration (Phillips and Moon (1999)). Alternatively, one can include non-stationary regressors in the model after first differencing them.

Another issue is binary regressors. Under Assumption 3.12, binary regressors are modeled with a linear probability model. In this case, a more suitable assumption could be a dynamic latent model. Another option could be a Markov chain assumption. This is left for further research.

3.1. Relationship between the baseline model and two-level panel data

Model 3.2 is related to a heterogeneous dynamic model for two-dimensional panel data under special conditions. To see this clearer, let us consider the following model:

$$y_{it} = \alpha_i + \rho_i y_{it-1} + x'_{it}\beta_{it} + \varepsilon_{it}, \qquad i = 1, 2, ..., N, \quad t = 1, 2, ..., T_i. \tag{3.5}$$

If the following assumptions hold, we can rewrite the two-level dynamic panel data model 3.5 as the three-level panel data model 3.2.

Assumption 3.13. *The individual additive unobserved effect is homogeneous within cluster*

$$\alpha_i = \alpha_g \quad \forall i \in g.$$

Under this assumption, the correlation of the additive unobserved individual heterogeneity with the regressors is equal within clusters. For instance, the innate ability of workers is equal within city. This is feasible if workers self-select into a city based on their ability.

Assumption 3.14. *The individual persistence parameter is homogeneous within cluster*

$$\rho_i = \rho_g \quad \forall i \in g.$$

This means that the persistence of the dynamic process is equal within clusters. An example of homogeneous persistence is equal consumption persistence within village. The homogeneity of consumption persistence within village could happen if village characteristics drive consumption habits.

Assumption 3.15. *The slope coefficients are conditional mean dependent on cluster belonging*

$$E[\beta_{it}|s_i^{(g)}, x_{i1}, x_{i2}, ..., x_{iT}] = \beta_g.$$

This assumption is equivalent to assumption 3.7.

The use of three-level or multi-dimensional panel data models surged due to the increasing availability of big data (Matyas (2017), Sarafidis and Wansbeek (2021)). The reason is that it allows to 1) control for unobserved heterogeneity that is not only individual and/or time specific (Sarafidis and Wansbeek (2021)), 2) accommodate belonging of each individual in clusters or groups (Sarafidis and Wansbeek (2021)), 3) deal with incidental parameter bias, 4) develop appropriate inference that considers the sampling uncertainty.

In this particular setting, three-level panel data allows the use of a mixed-coefficient structure with random coefficients within clusters and fixed coefficients across clusters, and the use of cluster additive specific effects instead of cluster-individual specific effects. Thus, we avoid the problem of incidental parameter bias. In addition, the available data $\{y_{it}\}_{i=1}^{N}$ can be partitioned into non-overlapping sub-samples reflecting the sampling design. Consequently, using three-level panel data permits to make explicit assumptions about cluster belonging, the stability of cluster belonging, and the relationship between clusters. Furthermore, we provide appropriate inference for two different sampling frameworks.

4. ESTIMATION

If we re-write model 1.1 using backward substitution, we obtain the following expression of the dependent regressor:

$$y_{git} = \rho_g^t y_{gi0} + \sum_{l=0}^{t} \rho_g^l (\alpha_{1,g} + x'_{git-l}(\beta_g + \lambda_{git-l})) + \sum_{l=0}^{t} (\rho_g^l) \varepsilon_{git-l}. \tag{4.6}$$

Using this result, the first lag of the dependent variable can be rewritten as:

$$y_{git-1} = \rho_g^{t-1} y_{gi0} + \sum_{l=0}^{t-1} \rho_g^l (\alpha_{1,g} + x'_{git-1-l}(\beta_g + \lambda_{git-1-l})) + \sum_{l=0}^{t-1} (\rho_g^l) \varepsilon_{git-1-l}. \tag{4.7}$$

It is easy to see from (4.7) that first-difference GMM estimation (Arellano and Bond (1991)) ignoring the clustering structure of the data leads to inconsistent estimates of the mean parameters. This is caused by the presence of the first lag and the cluster-specific effects in the right-hand side of the model causing endogeneity. Moreover, it is not possible to find an instrument that is uncorrelated with the composite error term and correlated with the regressors. [5]

Similarly, one could argue that the researcher could perform Mean Group estimation

[5] Ignoring cluster effects is equivalent to performing first-difference GMM estimation on the model: $\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x'_{it} \beta + \Delta u_{it}$ with: $\Delta u_{it} = \Delta y_{it-1} \alpha_{2,g} + \Delta x'_{it} \alpha_{3,g} + \Delta x'_{it} \lambda_{git} + \Delta \varepsilon_{it}$, $\alpha_{2,g} = \rho_g - E[\rho_g]$ and $\alpha_{3,g} = \beta_g - E[\beta_g]$. Thus, we would not have available instruments. Another possibility could be first-difference GMM estimation on the model in first differences using multiplicative cluster dummies when $T > 2$. But one could run into the problem of weak instrumental variables (Bun and Windmeijer, 2010).

per individual within cluster. Mean Group estimation could be used to estimate cluster-specific parameters only if the time dimension is bigger than the number of covariates and growing to infinity or using small sample debiasing techniques (Available only if $T > 3$). Thus, when the time dimension is fixed and the number of individuals per cluster is big it would be beneficial to use another estimation strategy.

In order to fill this gap, we propose a methodology that allows for the estimation of the mean cluster and the cluster-specific coefficients using a two-stage procedure. This estimation technique is an extension of the Mean-Group Estimator presented by Pesaran and Smith (1995). The two-stage procedure is the following:

First stage: In the first stage, one estimates the cluster-specific coefficients by exploiting the population moment condition for individual i within group g:

$$E[u_{git} z_{git}] = 0, \quad t = 1, 2, ..., T_{i_g}. \tag{4.8}$$

Moreover, the sample moment conditions per cluster g are given by:

$$\frac{1}{N_g} u_g' Z_g = 0, \quad g = 1, 2, ..., m. \tag{4.9}$$

It is easy to see that using the sample moment conditions 4.9 as estimating equations leads to a simple ordinary least squares estimator:

$$\hat{\theta}_{g,OLS} = (Z_g' Z_g)^{-1}(Z_g' y_g).$$

This estimator is not the most efficient since the model presents a non-i.i.d composite error term u_{git}. A straightforward solution is to set a GLS estimator:

$$\hat{\theta}_{g,GLS} = (Z_g' \Omega_g^{-1} Z_g)^{-1}(Z_g' \Omega_g^{-1} y_g),$$

where $\Omega_g = E[u_g u_g'] = diag(X_g)(I_{N_g T} \otimes \Delta_{\lambda_g}) diag(X_g) + \sigma^2_{\varepsilon_g} I_{N_g}$ if $T_g = T$. If $T_g \neq T$, one just needs to set up the adequate design matrix to allow unbalancedness in the time dimension.

Since Ω_g is unknown, we propose an estimation procedure for Ω_g in Subsection 4.1.

The assumptions of unobserved additive and multiplicative cluster fixed effects allow us to estimate the specific parameters by pooling observations within each cluster (Assumptions 3.5, 3.6, 3.7). Additionally, OLS or FGLS estimation is consistent under the assumptions presented in Section 3 because the model is dynamic complete conditional on cluster-specific effects. But the FGLS estimator is non-robust to violations of the assumptions 3.7 and 3.10 because the variance-covariance matrix is not diagonal if λ_{git} and ϵ_{git} are heteroskedastic, serially correlated and/or present cross-sectional correlation. In this case, it is better to use the OLS estimator with a fully robust variance estimator as explained in Section 7.

In the case of endogenous regressors, replacing the OLS or FGLS first-stage estimation with GMM estimation using instrumental variables is possible. In this case, identification is done using the population moment conditions $E[u_{git} p_{git}] = 0$ with p_{git} a vector of appropriate instruments. Moreover, for identification it is also necessary to assume that the number of instrumental variables is equal or larger than the endogenous regressors.

Second stage: the estimator of $E[\theta_g]$ is equal to the weighted average of the cluster estimated parameters. This is called the Mean Cluster estimator, and it is given by:

$$\hat{\hat{\theta}}_{MC} = \sum_{g}^{m} \hat{\pi}_g \hat{\theta}_g,$$

where $\hat{\pi}_g$ is an appropriate estimator of the importance of the cluster in the population, $\hat{\hat{\theta}}_{MC} = [\hat{\hat{\rho}} \quad \hat{\hat{\alpha}}_{1,g} \quad \hat{\hat{\beta}}]'$, $\hat{\theta}_g = [\hat{\rho}_g \quad \hat{\alpha}_{1,g} \quad \hat{\beta}_g]'$.
Under stratified sampling, we propose a weighted average of the cluster-specific coefficients where the weights represent the importance of each cluster in the population.

The difference between the Mean Cluster (MC) estimators and the Mean Group (MG) estimator proposed by Pesaran and Smith (1995) is that the MG is obtained by averaging the estimators for each individual in the panel. In contrast, the MC averages cluster pooled estimators.

The Mean Cluster estimator is also consistent under cluster sampling. Under cluster sampling, the proportion of observed clusters with respect to the total number of clusters in the population is lower than 1. Because the observed clusters are sampled with equal probability, we assign an equal weight ($\frac{1}{m}$) to all observed clusters. The assumptions for this setting are presented in Section 9 as well as the derivation of the statistical properties of the Mean Cluster estimator.

4.1. Variance-Covariance Estimation

In order to make GLS feasible, we propose a ridge regression estimation method of the variance-covariance components of $\triangle_{\lambda_g}$ and $\sigma^2_{\varepsilon_g}$.
First, we derive the linear decomposition of the variance-covariance matrix for each cluster:

$$\Omega_g = \sum_{k=1}^{K} \sum_{k'=1}^{K} \sigma_{\lambda_g,kk'} H_{g,kk',\lambda_g} + \sigma^2_{\epsilon_g} I_{n_g}. \tag{4.10}$$

with the design matrices equal to:

$$H_{g,kk',\lambda_g} = \tilde{X}_{g,k} \tilde{X}'_{g,k'},$$

where $\tilde{X}_{g,k} = diag(x_{git,k})$.
Now, we obtain a first stage estimator of the residuals for each cluster using OLS estimation $r_{g_{OLS}} = (I_{n_g} - Z_g(Z'_g Z_g)^{-1} Z'_g) y_g = M_g w_g$ where $Z_g \in \mathbb{R}^{n_g \times (K+1)}$ is the matrix stacking up all the observations for $z_{git} = [y_{git-1} \quad 1 \quad x'_{git}]'$. Then, it follows that:

$$E[r_{g_{OLS}} r'_{g_{OLS}} | X_g] = M_g \Omega_g M_g. \tag{4.11}$$

Replacing expression (4.10) into equation (12.58) and applying the vec operator, we obtain:

$$vec(E[r_{g_{OLS}} r'_{g_{OLS}} | X_g]) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \sigma_{\lambda_g,kk'} vec(M_g H_{g,kk',\lambda_g} M_g) + \sigma^2_{\epsilon_g} vec(M_g). \tag{4.12}$$

Now, we can rewrite the previous expression in matrix form:

$$vec(E[r_{g_{OLS}} r'_{g_{OLS}} | X_g]) = B_{\lambda_g} vec(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(M_g). \tag{4.13}$$

In order to avoid double estimation of the covariances in the variance-covariance matrix, we use the identity $vec(A) = Dvech(A)$ where A is a square symmetric matrix and we re-express the previous equation as:

$$vec(E[r_{g_{OLS}} r'_{g_{OLS}} | X_g]) = B_{\lambda_g} Dvech(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(M_g). \tag{4.14}$$

The expectation of the outer product of the residuals is replaced by the point estimator of the OLS residuals for each cluster and we add the error ν_g that captures the sampling error.

$$vec(r_{g_{OLS}} r'_{g_{OLS}}) = B_{\lambda_g} Dvech(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(M_g) + \nu_g. \tag{4.15}$$

Finally, notice that 12.62 is a simple linear model that can be rewritten as:

$$R_g = C_g \eta_g + \nu_g,$$

where:

$$\begin{aligned} R_g &= vec(r_{g_{OLS}} r'_{g_{OLS}}), \\ C_g &= [\quad B_{\lambda_g} D \quad vec(M_g)], \\ B_{\lambda_g} &= [vec(M_g H_{g,11,\lambda_g} M_g) \quad vec(M_g H_{g,12,\lambda_g} M_g) \quad ... \quad vec(M_g H_{g,KK,\lambda_g} M_g)], \\ \eta_g &= [vech(\triangle_{\lambda_g})' \quad \sigma^2_{\epsilon_g}]'. \end{aligned}$$

Now, the estimators of the elements of variance-covariance are obtained by minimizing the following penalized loss function:

$$L(\eta_g) = (R_g - C_g \eta_g)'(R_g - C_g \eta_g) + \tau \parallel \eta_g \parallel_2^2,$$

where τ is the penalisation parameter.

Notice, that for identification of η_g we implicitly assume:

Assumption 4.1. *$E[\nu_g C_g] = 0$.*

Assumption 4.1 states that the error term ν_g is orthogonal to the covariates included in C_g.

The penalization term using the l_2-norm allows us to tackle the problem of high multicollinearity in the matrix $C'_g C_g$. We follow Hoerl et al. (1975), Cule and De Iorio (2012) by estimating τ from the data as follows:

$$\hat{\tau} \geq \frac{\hat{\sigma}^2}{\hat{\beta}'_{OLS} \hat{\beta}_{OLS}},$$

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta}_{OLS})'(y - X\hat{\beta}_{OLS})}{NT - K - 1}$$

Following Hoerl and Kennard (1970), we can prove that the MSE of $\hat{\beta}_{FGLS}$ is monotonically decreasing on τ. Thus, we can choose a $\tau > 0$ that minimizes MSE. The choice $\hat{\tau}$ is heuristic, and we acknowledge that it might be possible to derive an optimal estimator of τ (this is left for further research).

C. Large and Huge Sample Size

When the sample size is big, there are problems due to memory requirements for storing vectorized matrices. In order to tackle this issue and reduce the computing requirements by half, we modify the method proposed above using the vech operator instead of the vec operator. It is possible to do this replacement since we are dealing with square symmetric matrices.

$$R_g = vech(r_g r_g'),$$
$$C_g = [B_{\lambda,g} \quad vech(M_g)],$$
$$B_{\lambda_g} = [vech(M_g H_{g,11,\lambda_g} M_g) \quad vech(M_g H_{g,12,\lambda_g} M_g) \quad ... \quad vech(M_g H_{g,KK,\lambda_g} M_g)].$$

This modification improves the computational performance but has limitations. For big samples, one needs computational algebra methods for matrix inversion and multiplication.

5. IDENTIFICATION

For identification, we can rewrite the model 3.2 as:

$$y_{git} = \rho_g y_{git-1} + \alpha_{1,g} + x'_{git}\beta_g + u_{git} = z'_{git}\theta_g + u_{git}, \tag{5.16}$$

where: $z_{git} = [y_{git-1} \quad 1 \quad x'_{git}]'$, $\theta_g = [\rho_g \quad \alpha_{1,g} \quad \beta'_g]'$, and $u_{git} = x'_{git}\lambda_{git} + \varepsilon_{git}$ is a composite error term.
Assumptions 3.8 and 3.9 imply the following orthogonality conditions: [6]

$$E[u_{git}x_{gis}] = 0, \quad s = 1, 2, ..., T_{i_g}, \quad i = 1, 2, ..., N_g, \quad g = 1, 2, ..., m, \tag{5.17}$$

$$E[u_{git}y_{git-1}] = 0, \quad t = 1, 2, ..., T_{i_g}, \quad i = 1, 2, ..., N_g, \quad g = 1, 2, ..., m. \tag{5.18}$$

Consequently, the moment conditions used for the estimation of the cluster-specific parameters are:

$$E[u_{git}z_{git}] = 0, \quad t = 1, 2, ..., T_{i_g}, \quad i = 1, 2, ..., N_g, \quad g = 1, 2, ..., m. \tag{5.19}$$

Note that we only use contemporaneous exogeneity for estimation of the cluster-specific parameters using cluster-specific data which is in line with Hsiao et al. (2019). According to Wooldridge (2010) contemporaneous exogeneity can be exploited when the variance-covariance of the model is diagonal as it is in each cluster.

Additionally, we also assume that the z_{git} is full rank which means that the regressors vary within cluster.

Assumption 5.1. *OLS: The matrix $E[z_{git}z'_{git}]$ is full rank.*

GLS: $E[u_g u'_g|Z_g]$ is positive definite and the matrix $E[Z'_g E[u_g u'_g]^{-1} Z_g] = Q_g$ is non-singular.

6. STATISTICAL PROPERTIES

In this section, we present the statistical properties of the cluster-specific estimators, the Mean Cluster estimator and the variance-covariance estimators using sequential asymp-

[6] According to Chamberlain (1987), the conditional moment $E[s|g(w)] = 0$ restriction implies that $E[g(w)s] = 0$ for any function $g(.)$ where s and w are two random variables.

totic theory with the number of individuals per cluster (N_g) growing to infinity and the time dimension (T_{i_g}) fixed. This implies that the total number of observations per cluster ($n_g = \sum_{i_g}^{N_g} T_{i_g}$) grows to infinity.

For convenience, we use the indexes i_g to refer to individual i belonging to cluster g and t_{i_g} for the time observation t of individual i_g.

6.1. *Cluster specific estimators*

THEOREM 6.1. *If i) Assumptions 3.1 to 3.15 and 5.1 hold, ii)* $\{y_{i_g}, x_{i_g}\}_{i_g=1}^{N_g}$ *is a sequence of random vectors containing* T_{i_g} *observations* $\forall g$*, iii)* $N_g \to \infty$ *and* T_{i_g} *fixed* ($n_g \to \infty$)*, then*

a) $\hat{\theta}_{g,GLS} \xrightarrow{p} \theta_g$, *b)* $\sqrt{n_g}(\hat{\theta}_g - \theta_g) \xrightarrow{d} N(0, Q_g)$.

where $Q_g = \underset{n_g\to\infty}{plim}(n_g^{-1} Z_g' \Omega_g^{-1} Z_g)^{-1}$.

6.2. *Variance Covariance Estimators*

THEOREM 6.2. *If i) Assumptions 3.1 to 3.15 and 5.1 hold, ii)* $\underset{n_g\to\infty}{plim} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} n_g^{-1} C_{i_g t_{i_g}} C'_{i_g t_{i_g}} = M_g$ *with* $||M_g||_F < \infty$*, iii)* $\nu_{i_g t_{i_g}} \sim iid(0, \sigma_\nu^2)$*, iv)* $\underset{n_g\to\infty}{lim} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} n_g^{-1} C_{i_g t_{i_g}} R_{i_g t_{i_g}} = 0$*, v)* $N_g \to \infty$ *and* T_{i_g} *fixed* ($n_g \to \infty$) *then*

a) $\hat{\Omega}_g \xrightarrow{p} \Omega_g$*, b)* $\sqrt{n_g}(\hat{\Omega}_g - \Omega_g) \xrightarrow{d} N(0, var(\hat{\Omega}_g))$.

6.3. *Mean Cluster Estimator*

THEOREM 6.3. *If i) Assumptions of theorems 6.1 and 6.2 hold* $\forall g$*, then*

$$\sqrt{N}(\hat{\hat{\theta}} - E[\theta_g]) \xrightarrow{d} N(0, Q),$$

where $Q = \sum_g \pi_g Q_g$.

7. MISSPECIFICATION OF THE VARIANCE-COVARIANCE MATRIX

As mentioned in Section 4, the FGLS estimator in step 1 is non-robust to violations of the assumptions 3.7 and 3.10 that state that λ_{git} and ϵ_{git} are not serially correlated and homoskedastic within cluster. The reason is that under contemporaneous exogeneity, the FGLS is consistent only if the variance-covariance matrix of the model is diagonal. When the variance-covariance matrix of the model is not diagonal, one requires the stronger condition of strict exogeneity of all regressors included in the model. But in model 3.2, the strict exogeneity of all right-hand side regressors does not hold because the lag of the dependent variable is present (Wooldridge (2010)). Then if λ_{git} and ϵ_{git} are serially correlated or/and heteroskedastic within cluster, it is better to estimate the cluster-specific parameters using OLS with a fully robust variance estimator.

The correlation of ϵ_{git} and/or λ_{git} within cluster could be caused due to sub-clustering. For instance, students within schools belong to the same village. In this situation, we can use a one-way sub-cluster fully robust variance estimator per cluster. If we index by j_g the sub-clusters in cluster g and we assume that there is no cross-correlation across sub-clusters, we can use the following within cluster one-way fully-robust variance estimator:

$$\widehat{Var(\hat{\beta}_g)} = (\sum_{j_g} X'_{j_g}\hat{\Omega}^{-1}_{j_g}X_{j_g})^{-1}(\sum_{j_g} X'_{j_g}\hat{\Omega}^{-1}_{j_g}\hat{u}_{j_g}\hat{u}'_{j_g}\hat{\Omega}^{-1}_{j_g}X_{j_g})(\sum_{j_g} X'_{j_g}\hat{\Omega}^{-1}_{j_g}X_{j_g})^{-1}.$$

The estimator is fully-robust for heteroskedasticity and serial-correlation within sub-cluster j_g using $m^{-1}_{j_g}\sum_{j_g}\hat{u}_{j_g}\hat{u}'_{j_g}$ as an estimator of $E[u_{j_g}u'_{j_g}]$. If the number of sub-clusters (m_{j_g}) grows and the number of observations within the sub-cluster is fixed, the Wald t-statistic is asymptotically normal (Wooldridge (2003), Cameron and Miller (2015)). If the number of sub-clusters is fixed, the sub-cluster robust variance-covariance estimator is downward-biased (Cameron and Miller (2015), Wooldridge (2003)) and the Wald t-statistic is no longer asymptotically normal distributed (Cameron and Miller (2015), Wooldridge (2003)). In this situation, the wild-cluster bootstrap-t method proposed by Cameron et al. (2008) cannot be used to provide valid inference for the Mean Cluster estimator. The reason is that the wild-cluster bootstrap-t is proposed for a homogeneous model. To provide valid inference for the Mean Cluster estimator proposed in this paper when there is within sub-cluster correlation and a small number of sub-clusters, there is a need to extend the one-way wild-cluster bootstrap-t method for the Mean Cluster estimator.

Another issue is that the one-way sub-cluster fully robust variance estimator is valid under the assumption that observations are not correlated across sub-clusters. A solution is using a two-way sub-cluster fully-robust variance estimator, but this estimator requires that the number of sub-clusters and the number of time observations per individual within cluster grow to infinity. If this is not the case, there is a need to extend the two-way wild-cluster bootstrap-t method for the Mean Cluster estimator.

8. RELATIONSHIP OF THE MEAN CLUSTER ESTIMATOR WITH OTHER ESTIMATORS

8.1. GLS estimation of a two-level panel data model with interactions of cluster dummies with the regressors

In this subsection, we compare the Mean Cluster estimator of model 3.2 with FGLS estimation of two-level panel data containing the interactions of cluster dummies ($s_i^{(g)}$) with the regressors in the model.

Corollary 8.1. *FGLS estimation of a model with interactions between the regressors and m cluster dummies is not equivalent to the Mean-Cluster FGLS estimator.*

To proof Corollary 8.1, we can re-write model as a two-level panel data model with interactions of m cluster dummies ($s_i^{(g)}$) with the regressors:

$$\begin{aligned} y_{it} = \sum_{g=1}^{m}\alpha_g s_i^{(g)} + \sum_{g=1}^{m}\rho_g y_{it-1}s_i^{(g)} + \\ \sum_{g=1}^{m} s_i^{(g)}x'_{it}\beta_g + x'_{it}\lambda_{it} + \epsilon_{it}. \end{aligned} \tag{8.20}$$

FGLS estimation of the whole system of equations is equivalent to the first step of the

Mean Cluster FGLS estimator if the random terms ϵ_{it} and λ_{it} are not correlated across clusters (Assumption 3.7 and Assumption 3.15 hold). If assumptions 3.7 and 3.15 do not hold, the first step of the Mean Cluster estimator is less efficient than FGLS estimation on the whole system of equations (Greene (2008)).

In addition, we decompose the heterogeneous parameters as a sum of their overall means and the deviations from their cluster means as follows:

$$\alpha_g = E[\alpha_g] + (\alpha_g - E[\alpha_g]), \tag{8.21}$$

$$\rho_g = E[\rho_g] + (\rho_g - E[\rho_g]), \tag{8.22}$$

$$\beta_{git} = E[\beta_{git}] + (\beta_g + \lambda_{it} - E[\beta_{git}]). \tag{8.23}$$

Using these new expressions to re-write the three-level panel data 3.2 as a two-level panel data model, we obtain:

$$\begin{aligned} y_{it} = E[\alpha_g] + \sum_{g=1}^{m}(\alpha_g - E[\alpha_g])s_i^{(g)} + E[\rho_g]y_{it-1} + \sum_{g=1}^{m}(\rho_g - E[\rho_g])y_{it-1}s_i^{(g)} + x'_{it}E[\beta_{git}] + \\ \sum_{g=1}^{m} s_i^{(g)} x'_{it}(\beta_g - E[\beta_{git}]) + x'_{it}\lambda_{it} + \epsilon_{it}. \end{aligned} \tag{8.24}$$

Model 8.24 can be expressed in matrix form as follows:

$$\begin{aligned} y = \iota_n E[\alpha_g] + S(\alpha - \iota_m E[\alpha_g]) + E[\rho_g]y_{-1} + y_{1,-1}(\rho - \iota_m E[\rho_g]) + XE[\beta_{git}] \\ + X_1(\beta - \iota_m \otimes E[\beta_{git}]) + X_2\lambda + \epsilon. \end{aligned} \tag{8.25}$$

where y is a vector collecting the observations of all the clusters ($n = \sum_g^m n_g$), S is a matrix of m dummy variables with dimensions $n \times m$, y_{-1} is a vector collecting the first lag of the dependent variable with dimensions $n \times 1$, $y_{1,-1}$ is a block-diagonal matrix containing all the cluster-specific lagged variables $y_{g,-1}$ with dimensions $n \times m$, X is a matrix collecting K regressors with dimensions $n \times K$, $X_1 = diag(X_g)$ is a block-diagonal matrix that collects the group-specific matrices of regressors (X_g) with dimensions $n \times Km$, $X_2 = diag(x'_{it})$ with x'_{it} in the diagonal with dimensions $n \times Kn$. In addition, ι_n is a vector of ones with dimensions $n \times 1$, ι_m is a vector of ones with dimensions $m \times 1$, $(\alpha - \iota_m E[\alpha_g])$ is a vector of dimensions $mx1$, β is a vector containing all the m cluster slope parameters with dimensions $mK \times 1$, λ is a vector with dimensions $nK \times 1$ stacking up λ_{it}.

If we assume that the deviations of the parameters from their cluster means are fixed, we cannot estimate model 8.25. The reason is that the matrix containing all regressors of model 8.25 is rank deficient as a consequence of the perfect linear relationship between ι and S as well as between X and X_1. In other words, there is perfect multicollinearity in the model. For consistent estimation of model 8.24, we need to impose the following restrictions:

$$\sum_{g=1}^{m}(\alpha_g - E[\alpha_g]) = 0, \tag{8.26}$$

$$\sum_{g=1}^{m}(\rho_g - E[\rho_g]) = 0, \tag{8.27}$$

$$\sum_{g=1}^{m}(\beta_g - E[\beta_{git}]) = 0. \tag{8.28}$$

In conclusion, the Mean Cluster estimator can be interpreted as a method that imposes these restrictions through a two-step procedure. In addition, the Mean Cluster estimator is the most efficient estimator if the clusters are uncorrelated with each other. If the clusters are inter-correlated, the best estimation procedure is a pooled estimation of model 8.24 imposing the restrictions 8.26 to 8.28.

Corollary 8.2. *FGLS estimation of a model with interactions between the regressors and $m-1$ dummies is not equivalent to the Mean-Cluster FGLS estimator.*

Corollary 8.2 states that estimation of model 8.25 including $m-1$ dummies and their interactions with the regressors does not retrieve the average partial effects. To see this clearer, consider the following modified model:

$$y = \iota_n\alpha_m + \tilde{S}\psi_1 + \rho_m y_{-1} + \tilde{y}_{1,-1}\psi_2 + X\beta_m + \tilde{X}_1\psi_3 + X_2\lambda + \epsilon. \tag{8.29}$$

where $\tilde{S}$ contains the first $m-1$ columns of S, $\tilde{y}_{1,-1}$ contains the first $m-1$ columns of matrix $y_{1,-1}$ and $\tilde{X}_1$ contains the first $(m-1)K$ columns of the matrix X. This means that the baseline category is the cluster m.
It is clear from model 8.25 that the estimated coefficients of the regressors are just the cluster-specific coefficients for cluster m. The estimated coefficients of the interaction terms of the regressors with the $m-1$ dummies (ψ_2, ψ_3) are the deviations of their parameters from the coefficient of the baseline cluster. In order to obtain the average partial effects, one needs to take a four-step procedure to retrieve the cluster-specific parameters. The estimation procedure of the average partial effects using model 8.25 is the following:

Step 1: Estimate model 8.25 using pooled FGLS.

Step 2: Obtain the $m-1$ cluster-specific estimated parameters by adding their deviations to the cluster-specific parameter m. For instance, the estimated intercept of cluster j is equal to $\hat{\alpha}_j = \hat{\alpha}_m + \hat{\psi}_{1,j}$.

Step 3: Obtain the standard errors of the cluster-specific parameters. For instance, $s.e.(\hat{\alpha}_j) = \sqrt{var(\hat{\alpha}_m) + var(\hat{\psi}_{1,j}) + 2cov(\hat{\alpha}_m, \hat{\psi}_{1,j})}$.

Step 4: Obtain the average of the cluster-specific parameters.

In conclusion, the Mean Cluster estimator is a straightforward procedure to estimate the average partial effects. It is the most efficient estimator if the clusters are not correlated across each other. But if the clusters are correlated across each other, one needs to estimate 8.25 using the procedure described above with FGLS.

8.2. *First-difference GMM estimator*

Corollary 8.3. *The first-difference GMM (Arellano and Bond (1991)) estimator of model 3.2 is an inconsistent estimator of $E[\rho_g]$ and $E[\beta_{git}]$ and it is equal to a weighted averaged of the cluster-specific parameters.*

Proof. Taking the first-difference of model 3.2 and using the deviations of the cluster-specific parameters from their overall means (8.21, 8.22 and 8.23) leads to the following model:

$$\Delta y_{git} = E[\rho_g]\Delta y_{git-1}+(\rho_g-E[\rho_g])\Delta y_{it-1}+\Delta x'_{git}E[\beta_{git}]+\Delta(x'_{git}(\beta_{git}-E[\beta_{git}]))+\Delta\epsilon_{git}. \qquad (8.30)$$

If we stack up the cluster-specific first-differenced observations, we obtain:

$$\begin{bmatrix}\Delta y_1\\ \Delta y_2\\ ...\\ \Delta y_m\end{bmatrix} = E[\rho_g]\begin{bmatrix}\Delta y_{1,-1}\\ \Delta y_{2,-1}\\ ...\\ \Delta y_{m,-1}\end{bmatrix}+\begin{bmatrix}\Delta X_1\\ \Delta X_2\\ ...\\ \Delta X_m\end{bmatrix}E[\beta_{git}]+\begin{bmatrix}\widetilde{\Delta y}_{1,-1}\\ \widetilde{\Delta y}_{2,-1}\\ ...\\ \widetilde{\Delta y}_{m,-1}\end{bmatrix}+\begin{bmatrix}\widetilde{\Delta X}_1\\ \widetilde{\Delta X}_2\\ ...\\ \widetilde{\Delta X}_m\end{bmatrix}+\begin{bmatrix}\Delta \varepsilon_1\\ \Delta \varepsilon_2\\ ...\\ \Delta \varepsilon_m\end{bmatrix}. \qquad (8.31)$$

with $\widetilde{\Delta X}_g$ stacking up $\Delta(x'_{git}(\beta_{git}-E[\beta_{git}]))$, and $\widetilde{\Delta y}_{g,-1}$ stacking up $(\rho_g-E[\rho_g])\Delta y_{git-1}$.

If we collect the regressors $\Delta y_{g,-1}$ and ΔX_g in a matrix ΔZ_g, the parameters of interest $E[\rho_g]$ and $E[\beta_{git}]$ in the column vector $\tilde{\theta}$, and the random components $\widetilde{\Delta y}_{g,-1}$, $\widetilde{\Delta X}_g$, $\Delta\varepsilon_g$ in a composite error term $\Delta\tilde{u}_g$ we obtain:

$$\begin{bmatrix}\Delta y_1\\ \Delta y_2\\ ...\\ \Delta y_m\end{bmatrix} = \begin{bmatrix}\Delta Z_1\\ \Delta Z_2,\\ ...\\ \Delta Z_m\end{bmatrix}\tilde{\theta}+\begin{bmatrix}\Delta \tilde{u}_1\\ \Delta \tilde{u}_2\\ ...\\ \Delta \tilde{u}_m\end{bmatrix}. \qquad (8.32)$$

Then, using as instrumental variables $y_{git-s}, \forall s > 1$, x_{gi} and calling the matrix of instruments $W = diag(W_g)$ we get that the first-difference GMM estimator is equal to:

$$\hat{\tilde{\theta}}_{F-GMM} = (\Delta Z'WV^{-1}W'\Delta Z)^{-1}(\Delta Z'WV^{-1}W'\Delta y),$$

with $V = E[\Delta\tilde{u}\Delta\tilde{u}'] = diag(E[\Delta\tilde{u}_g\Delta\tilde{u}'_g]) = diag(V_g)$.

Then, the first-difference GMM estimator ($\hat{\tilde{\theta}}_{F-GMM}$) can be re-written as follows:

$$\hat{\tilde{\theta}}_{F-GMM} = \tilde{\theta} + (\Delta Z'WV^{-1}W'\Delta Z)^{-1}(\Delta Z'WV^{-1}W'\Delta\tilde{u}),$$

$$plim\hat{\tilde{\theta}}_{F-GMM} = \tilde{\theta} + (plim\frac{1}{n}\Delta Z'WV^{-1}W'\Delta Z)^{-1}(plim\frac{1}{n}\Delta Z'WV^{-1}W'\Delta\tilde{u}).$$

We can see that the last term does not vanish because of the presence of $\Delta y_{-1}(\rho_g - E[\rho_g])$ in the composite error term $\Delta\tilde{u}$.

In addition, we can re-write the first-difference GMM estimator as the weighted sum of the cluster-specific first-difference GMM estimators as follows:

$$\hat{\hat{\theta}}_{F-GMM} = (\sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g)^{-1} (\sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta y_g),$$

$$\hat{\hat{\theta}}_{F-GMM} = (\sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g)^{-1} (\sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g \hat{\theta}_{g,F-GMM}).$$

Applying the plim operator to the last expression gives:

$$plim\hat{\hat{\theta}}_{F-GMM} = (plim\frac{1}{n} \sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g)^{-1} (plim\frac{1}{n} \sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g \hat{\theta}_{g,F-GMM}),$$

$$plim\hat{\hat{\theta}}_{F-GMM} = \sum_g w_g A^{-1} b_g \theta_g,$$

with $A = {}_{n \to \infty}^{plim} \frac{1}{n} \sum_g \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g$, $b_g = {}_{n_g \to \infty}^{plim} \frac{1}{n_g} \Delta Z_g' W_g V_g^{-1} W_g' \Delta Z_g$, and $w_g = lim\frac{n_g}{n}$.

The reason for the inconsistency of the first-difference GMM estimator is that ignoring heterogeneity in the coefficients causes endogeneity and there are no available instrumental variables no matter how long is the time dimension. In contrast, the Mean-Cluster estimator is a consistent estimator of the average partial effects.

8.3. Mean-Group estimator

Corollary 8.4. *When the time dimension is equal to 3, the Mean-Group estimator is unfeasible.*

Corollary 8.5. *When the time dimension is long, the Mean-Cluster estimator and the Mean-Group estimator are consistent estimators of the average partial effects $E[\theta_g]$ of model 3.2 under stratified sampling.*

Proof. The Mean-group estimator is equal to:

$$\hat{\theta}_{MG} = \frac{1}{N} \sum_g \sum_{i_g} \hat{\theta}_{i_g,OLS}, \tag{8.33}$$

with $\hat{\theta}_{i_g,OLS} = (Z_{i_g}' Z_{i_g})^{-1} (Z_{i_g}' y_{i_g})$.
Now, we have that:

$$plim\hat{\theta}_{MG} = \frac{1}{N} \sum_g \sum_{i_g} plim\hat{\theta}_{i_g,OLS}, \tag{8.34}$$

$$plim\hat{\theta}_{MG} = \sum_g \pi_g \theta_g = E[\theta_{git}], \tag{8.35}$$

because:

$$plim\hat{\theta}_{i_g,OLS} = (plim\frac{1}{n_g} Z_{i_g}' Z_{i_g})^{-1} (plim\frac{1}{n_g} Z_{i_g}' y_{i_g}),$$

$$plim\hat{\theta}_{i_g,OLS} = \theta_g + (plim\frac{1}{n_g}Z'_{i_g}Z_{i_g})^{-1}(plim\frac{1}{n_g}Z'_{i_g}\varepsilon_{i_g}),$$

$$plim\hat{\theta}_{i_g,OLS} = \theta_g.$$

9. CLUSTER SAMPLING

In this section, we present the assumptions under cluster sampling. The identification strategy proposed in section 5 is still valid while the estimation strategy proposed in section 4 needs to be modified by using weights equal to $1/m$. In addition, the statistical properties of the Mean Cluster estimator are different from the ones presented in section 6.

9.1. Assumptions

For the cluster sampling setting, we replace assumptions 3.1, 3.2, and 3.3 by the following ones:

Assumption 9.1. *The proportion of observed clusters q is lower than 1.*

This assumption means that the number of observed clusters m is not equal to the total number of clusters in the population.

Assumption 9.2. *Cluster size is homogeneous* $N_g = N^\kappa$ *with* $\kappa \in (0,1)$ *and* $T_{i_g} = T$.

Under this assumption, the number of observations per cluster is equal and we are back in a balanced panel data setting. This assumption is done for convenience to derive the asymptotic distribution of the Mean Cluster estimator under cluster sampling. An extension to a setting with an unbalanced cluster size under cluster sampling is left for further research.

Assumption 9.3. *Clusters are asymptotically negligible.*

$$\underset{N\to\infty}{lim}\frac{N_g}{N} = 0.$$

This assumption is in line with Hansen and Lee (2019) and it allows the number of individuals within clusters to grow with the total number of individuals but not at a proportional rate.

Assumption 9.4. *The number of clusters grows at a slower rate than the square of the total number of observations in the cluster.*

$$m \text{ is a monotonic function of } n_g,\ \frac{\sqrt{m}}{n_g} \to 0 \text{ as } n_g \to \infty.$$

This condition is necessary to guarantee the consistency of the Mean Cluster estimator.

Assumption 9.5. *Random cluster additive specific effects $\alpha_{1,g}$.*

$$E[\alpha_{1,g}] = 0$$

$$E[\alpha_{1,g}^2] = \sigma_{\alpha_1}^2.$$

Assumption 9.6. *Random cluster specific persistence parameter.*

$$\rho_g \in (-1, 1),$$

with $\alpha_{2,g} = \rho_g - E[\rho_g]$ and $E[\alpha_{2,g}^2] = \sigma_{\alpha_2}^2$.

Assumption 9.7. *Random cluster-individual-time specific coefficients.*

$$\beta_{git} = \beta_g + \lambda_{git},$$

$$E[\lambda_{git}\lambda'_{g'i't'}] = \begin{cases} \Delta_{\lambda_g} & \text{if} \quad g = g', i = i' and t = t', \\ 0 & \text{otherwise.} \end{cases}$$

with $\alpha_{3,g} = \beta_g - E[\beta_g]$ and $E[\alpha_{3,g}\alpha'_{3,g'}] = \Delta_{\alpha_3}$ if $g = g'$ and 0 otherwise.

The unobserved coefficient vector is composed of a cluster-specific coefficient vector (β_g) and a heteroskedastic random component (λ_{git}) that captures the multiplicative heterogeneity over time for each individual of cluster g.

Assumption 9.8. *Cluster specific effects have non-zero mean conditional on covariates.*

$$E(\alpha_g | x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}) \neq 0,$$

with $\alpha_g = [\alpha_{1,g} \quad \alpha_{2,g} \quad \alpha'_{3,g}]'$.

We could also assume that $E[\alpha_{1,g} z_{git}] \neq \mathbf{0}$, $E[\alpha_{2,g} y_{git-1}] \neq 0$, $E[\alpha'_{3,g} x_{git}] \neq 0$. The latter means that regressors are freely correlated to cluster unobserved heterogeneity.

Assumption 9.9. *The random cluster-individual-time effects have zero mean conditional on the covariates and cluster effects.*

$$E[\lambda_{git} | x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}, \alpha_g] = 0.$$

Assumption 9.10. *Strict exogeneity of the covariates with the disturbance term conditional con cluster effects.*

$$E[\varepsilon_{git} | x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}, \alpha_g] = 0.$$

9.2. Statistical properties of the Mean Cluster Estimator under cluster sampling

Theorem 9.1. *If i) assumptions 3.10 to 3.15 and 9.1 to 9.10 hold, ii) $N_g = N^\kappa$ with $\kappa \in (0, 1)$, iii) $\hat{\bar{\theta}}$ is consistent and asymptotically normal.*

$$\sqrt{m}(\hat{\hat{\theta}} - E[\theta_g]) \xrightarrow{d} N(0, \triangle_\alpha),$$

$$\textit{with } \triangle_\alpha = \begin{bmatrix} \sigma^2_{\alpha_1} & 0 & 0 \\ 0 & \sigma^2_{\alpha_2} & 0 \\ 0 & 0 & \Delta_{\alpha_3} \end{bmatrix}.$$

Condition ii) states that clusters have a homogeneous number of individuals. Under this condition, the Mean Cluster estimator is $\sqrt{m}$-consistent. The derivation of the asymptotic distribution of the Mean Cluster estimator under cluster sampling with heterogeneous cluster sizes is left for further research.

10. ARE CLUSTER EFFECTS ENOUGH?

The estimator proposed in subsection 4 can have two potential biases: i) incidental parameter bias, and ii) misspecification bias.

The incidental parameter bias occurs when the number of observations per group n_g is small, which happens when the number of individuals per cluster is small. A solution for this issue is debiasing.

The estimator is also subject to misspecification bias if the assumption $E[\lambda_{git}|x_{gi}, y_{git-1}] = 0$ fails. This happens when fixed cluster effects are not enough to account for possible correlated residual cluster-individual specific unobserved heterogeneity. Another possible source of misspecification bias occurs when the assumed coefficient structure is not correct. We can see that $\beta_{git} = \beta_{gi} + \lambda_{gt}$ is also a plausible structure. In this case, $\beta_{git} = \beta_g + \lambda_{git}$ is not the correct specification. In order to address these issues, we present specification tests in the following section. we also present an extension of model 1.1 that includes cluster-individual additive effects.

Finally, we abstract from misspecification bias caused by an incorrect clustering structure because we assume that clustering is known. This is possible when the available sample is drawn from a population that is divided into well-known clusters such as a country and its municipalities. Examples of these type of data are longitudinal data for households, and firm-employee matched data. The clustering assumption used in this paper is different from the one presented by Bester and Hansen (2016).

11. SPECIFICATION TESTS

In order to test the assumption of clustered heterogeneity, we propose two specification tests that are extensions of the Hausman test (Hausman and Taylor, 1981).

First, testing the null hypothesis of cluster additive and multiplicative heterogeneity versus cluster-individual additive and multiplicative heterogeneity is not feasible when the time dimension is as short as 3.

Second, testing the null hypothesis of complete homogeneity versus cluster additive and multiplicative heterogeneity is possible. In this case, we propose to compare the Mean Cluster estimator with the Pooled OLS estimator. More specifically, the null and alternative hypothesis are the following:

H_o : $\hat{\beta}_{MC}$ consistent and inefficient, $\hat{\beta}_{POLS}$ consistent and efficient.

H_1 : $\hat{\beta}_{POLS}$ inconsistent and $\hat{\beta}_{MC}$ consistent and most efficient.

The statistic is given by:

$$Q = (\hat{\beta}_{MC} - \hat{\beta}_{POLS})' Var(\hat{\beta}_{MC} - \hat{\beta}_{POLS})^{-1} (\hat{\beta}_{MC} - \hat{\beta}_{POLS}),$$

follows a $\chi^2_{df=K}$.

In addition, testing the null hypothesis of cluster additive and multiplicative effects versus cluster-individual additive and cluster multiplicative heterogeneity is also viable. In this case, we propose to use a Hausman-type test that compares Mean Cluster estimators vs. a Mean Cluster First-difference GMM estimator or the Mean Cluster estimator using a Mundlak approach.

The study of the statistical properties of these tests is left for further research.

12. RELAXING THE ASSUMPTION OF CLUSTER ADDITIVE SPECIFIC EFFECTS

12.1. Initial conditions generated from the stationary distribution

In this subsection, we relax the assumption of additive cluster specific effects and allow for the presence of additive cluster-individual correlated random effects. Therefore, Assumption 3.5 is replaced by the following one:

Assumption 12.1. *Correlated cluster-individual additive specific random effects* $\alpha_{1,gi}$.

The inclusion of cluster-individual additive effects allows to control for endogeneity of the regressors that might not be captured by the cluster additive fixed effects. In particular, we consider the following extension of the model 1.1:

$$y_{git} = \alpha_{1,gi} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}, \tag{12.36}$$

where $\alpha_{1,gi}$ is a cluster-individual specific correlated random effect.

The estimation of model 12.36 with short time dimension has two main problems: i) the incidental parameter bias caused by the presence of the cluster-individual specific effects and ii) the impact of unobserved initial values (y_{gi0}) on the estimation.

In order to deal with the incidental parameter bias, we use a mean conditional approach instead of a linear difference approach. We choose the mean conditional approach because it is appropriate for heterogenous dynamic panel data models. As explained by Hsiao (2020), in this approach it is needed to use a linear approximation of $E(\alpha_{gi}|x_{it})$ to model the correlation of the regressors with the cluster-individual unobserved effects (This was a suggestion of Mundlak (1961) and Chamberlain (1979)). Following this suggestion, we re-express $\alpha_{1,gi}$ as a linear projection on the individual means of the regressors:

$$\alpha_{1,gi} = \bar{x}'_{gi.}\varphi_g + \upsilon_{gi}, \tag{12.37}$$

where $\bar{x}_{gi.} = T^{-1}\sum_{t=1}^{T} x_{git}$, υ_{gi} is an orthogonal error term such that $E(\upsilon_{gi}|\bar{x}_{gi.}) = 0$, and φ_g is a vector of unobserved parameters.

This linear projection can be replaced in model 12.36 obtaining:

$$y_{git} = \bar{x}'_{gi.}\varphi_g + \rho_g y_{git-1} + x'_{git}\beta_{git} + \upsilon_{gi} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}. \tag{12.38}$$

Now, it is only left the problem of unobserved initial conditions dependency. Modifying Assumption 3.11 to allow for the presence of cluster-individual additive effects yields:

$$y_{gi0} = \rho_g^{h_{i_g}} y_{gi,-h_{i_g}} + \alpha_{1,gi}\frac{1-\rho_g^{h_{i_g}}}{1-\rho_g} + \sum_{l=0}^{h_{g_i}} \rho_g^l x'_{gi-l}\beta_{gi-l} + \sum_{l=0}^{h_{i_g}} \rho_g^l \varepsilon_{gi-l}. \tag{12.39}$$

If we assume that $h_{g_i} \to \infty$, we can re-write the initial conditions as follows:

$$y_{gi0} = \frac{\alpha_{gi}}{1-\rho_g} + \sum_{l=0}^{\infty} \rho_g^l x'_{gi-l}\beta_{gi-l} + \sum_{l=0}^{\infty} \rho_g^l \varepsilon_{gi-l}. \tag{12.40}$$

Following Hsiao (2020), we re-call the terms of equation 12.40 such that the equation of the initial values is:

$$y_{gi0} = \frac{\alpha_{gi}}{1-\rho_g} + \psi_{gi0} + \varepsilon_{0i}. \tag{12.41}$$

Replacing the linear projection of the individual effects on the individual mean of the regressors to obtain:

$$y_{gi0} = \frac{\bar{x}'_{gi.}\varphi_g}{1-\rho_g} + \psi_{gi0} + \frac{\upsilon_{gi}}{1-\rho_g} + \varepsilon_{0i}. \tag{12.42}$$

In this equation, it is clear that we still have the problem of incidental parameters due to the presence of ψ_{gi0}. In order to deal with this issue, we follow Hsiao (2020) and assume that $E(\psi_{gi0}|x_{gi}) = \bar{x}'_{gi}\phi^*_g$. This is possible under Assumptions 3.11 and 3.12.

The combination of 12.38, 12.42, and $E(\psi_{gi0}|x_{gi}) = \bar{x}'_{gi}\phi^*_g$ leads to the system of equations:

$$\begin{aligned} y_{git} &= \bar{x}'_{gi.}\varphi_g + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon^*_{git}, \quad t = 1, ..., T_{i_g}, \\ y_{gi0} &= \frac{\bar{x}'_{gi.}\varphi_g}{1-\rho_g} + \bar{x}'_{gi}\phi^*_g + \frac{\upsilon_{gi}}{1-\rho_g} + \varepsilon_{0i}. \end{aligned} \tag{12.43}$$

where $\varepsilon^*_{git} = \varepsilon_{git} + \upsilon_{gi}$.

For estimation of the system 12.43, we propose two different methodologies. The first one is a Bayesian hierarchical estimator with a prior for the initial conditions. The second one proposes to estimate the equation of the dependent variable conditional on the initial conditions. These methods are described in the following subsections:

12.1.1. Bayesian estimation In order to set up the Bayesian estimator, we define the likelihood of the observed data by:

$$L_{\zeta|y,y_{-1},X} = \prod_{g}^{m}\prod_{i}^{N_g} L(\zeta_g|y_{gi}, X_{gi}), \tag{12.44}$$

where $\zeta_g = [\rho_g \quad \beta_g \quad \phi_g \quad \sigma^2_{\varepsilon^*}]'$, $\zeta = [\zeta_1 \quad \zeta_2 \quad ... \quad \zeta_m]'$, $L(\zeta_g|y_{gi}, X_{gi}) = f(y_{gi}|X_{gi};\zeta_g)$ with $f(y_{gi}|X_{gi};\zeta_g)$ representing the multivariate normal distribution with variance equal to $\sigma^2_\varepsilon I_T + \sigma^2_\upsilon \iota_T \iota'_T$ and with expectation equal to $\mu_{y,gi} = \rho_g y_{gi-1} + diag(x_{gi})\beta_{gi} + \iota_{T_{i_g}}\bar{x}'_{gi.}\varphi_g$. The prior distributions for the fixed parameters are:

$$(\beta_g|\beta) \sim N(\beta, H_{\Delta_{\alpha,2}} H'_{\Delta_{\alpha,2}}),$$

$$(\rho_g|\rho) \sim N(\rho, \sigma^2_\rho),$$

$$(\varphi_g|\varphi) \sim N(\varphi, FF').$$

While the prior for the random effects is:

$$\lambda_{git} \sim N(0, H_{\Delta_\lambda} H'_{\Delta_\lambda}),$$

$$H_{\Delta_\lambda} \sim LKJ(2).$$

The prior distribution of the variance σ^2_ε is half-normal with a location parameter equal to 0.5 and a scale parameter equal to 0.2. The prior distribution of the lower triangular matrix H_{Δ_λ} is Lewandowski-Kurowicka-Joe (LKJ) with parameter equal to 2. The value of the parameter of the LKJ prior means that the matrix has a low correlation.

Notice that the prior set-up imposes a non-centered parametrization on β_{git} such that:

$$\beta_{git} = \beta_g + H_{\Delta_\lambda} z_{git}, \tag{12.45}$$

where z_{git} is a standard multivariate normal variable and H_{Δ_λ} is the Cholesky factor of the variance-covariance matrix of λ_{git}.

This non-centered parameterization improves the convergence of the Hamiltonian Monte Carlo (HMC) algorithm because it reduces the correlation of the parameters (Frühwirth-Schnatter and Tüchler, 2008; Betancourt and Girolami, 2013). This reduction of the correlation permits the exploration of the whole parameter space improving the mixing of the chains.

Remark 12.1. According to Rossi and Allenby (2009) and Rendon (2013), imposing prior distributions only for the parameters of the model leads to a fixed effects specification. Thus, there is not any prior specification for the hyper-parameters of the priors. Therefore, a Bayesian model for a fixed effects specification has only first-stage priors while a Bayesian model for a random effects specification includes second-stage or hyper-priors.

Under the simplifying assumption that y_{gi0} is known, we could just set up a naive Bayesian estimator. But the assumption that y_{gi0} is fixed is not plausible. Its failure leads to inconsistent estimates. This is why, we relax it and set up the following prior distribution for the initial conditions:

$$f_{y_{gi0}} \sim N(\mu_{0,gi}, \sigma^2_{y_0}), \tag{12.46}$$

where $\mu_{0,gi} = \frac{\alpha_{gi}}{1-\rho_g} + \bar{x}'_{gi}\phi^*_g$, and the prior distribution of the variance $\sigma^2_{y_0}$ is half-normal with location parameter equal to 0.5 and scale parameter equal to 0.2.

Remark 12.2. Assuming that y_{gi0} comes from the stationary distribution means that the initialization of the process happened a long time ago ($h_{i_g} \to \infty$). This implies that the parameter b_g is equal to 0.

12.1.2. Conditioning on the initial value Another option for consistent estimation of the parameters of interest is the estimation of model 12.36 after conditioning on the initial value. For this purpose, we follow Hsiao (2020) and condition the first equation of the system 12.43 on the initial value y_{gi0} leading to:

$$y_{git} = \bar{x}'_{gi.}\varphi^*_g + \rho_g y_{git-1} + x'_{git}\beta_{git} + y_{gi0}\tilde{b}_g + \varepsilon^*_{git}, \quad t = 1, ..., T_{i_g} \tag{12.47}$$

Estimation of this model can be done using the Mean Cluster estimator with FGLS in the first stage.

12.2. Initial conditions not generated from the stationary distribution

A failure of the assumptions that the DGP of y_{gi0} is generated from the stationary distribution (Assumption 3.11), and the stationary regressors (Assumption 3.12) renders the Bayesian estimator presented in the previous section inconsistent. In order to relax the assumptions 3.11 and 3.12, we assume that y_{gi0} is unknown and that it does not come from the stationary distribution. This is done in order to avoid making assumptions regarding the exogenous regressors. As explained by Heckman (1987) and stated in Assumption 3.12, we need to make assumptions about the stationarity of the explanatory regressors and rule out time and age trends when the initial conditions are generated from the stationary process.

12.2.1. Bayesian Estimation In order to propose a prior that does assume that initial conditions (y_{gi0}) are generated from the stationary distribution, we propose the following joint prior:

$$\begin{pmatrix} y_0 \\ \\ \theta \end{pmatrix} \sim N \begin{pmatrix} \iota_{mNT} \otimes \mu_y & \\ & , \ \Sigma_{y,\theta} \\ \iota_{mNT} \otimes \bar{\theta} & \end{pmatrix},$$

with:

$$\Sigma_{y,\theta} = \begin{pmatrix} \sigma^2_{y_0} I_{mN} & \Sigma_{y_0,\beta} & \sigma_{y_0,\rho} \\ \Sigma_{y_0,\beta} & \Sigma_\beta & 0 \\ \sigma_{y_0,\rho} & 0 & \Sigma_\rho \end{pmatrix}.$$

A similar idea was presented by Sims (2000) and Heckman (1987). They defined a joint prior for the initial conditions and the coefficient vector.

Implementation of this Bayesian estimator is not straightforward due to the correlations between the initial conditions and the parameters of interest and the unknown initial values and we leave it for further research.

12.2.2. Conditioning on the initial value Another approach to deal with the problem of initial conditions dependency without making the restrictive assumptions of initial values generated from the stationary distribution and stationary regressors (Assumptions 3.11 and 3.12) is to project the cluster-individual specific effects into the column space of the regressors x_{git} and the initial condition y_{gi0} as proposed by Wooldridge (2005b) and the individual average of the regressors (Mundlak approach Mundlak (1978)). More specifically, the linear projection of the cluster-individual additive effects is equal to:

$$\alpha_{gi} = \bar{x}'_{gi}b^*_g + y_{gi0}a_g + \tilde{\omega}_{gi}, \tag{12.48}$$

Replacing this expression into model 12.36 leads to:

$$y_{git} = \bar{x}'_{gi}b^*_g + y_{gi0}a_g + \rho_g y_{git-1} + x'_{git}\beta_{git} + \tilde{\omega}_{gi} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}. \tag{12.49}$$

In addition, we can project the initial conditions on the individual averages of the regressors following ... :

$$y_{gi0} = \bar{x}'_{gi}c^*_g + \varepsilon_{gi0}, \tag{12.50}$$

Stacking up the initial conditions and the augmented model, we can estimate model 12.47 using the Mean-Cluster estimator with a factor approach in the first stage following Bai (2013). For this purpose, re-write the cluster specific model as follows:

$$B_g\tilde{y}_{gi} = d_{1gi}\bar{x}'_{gi}b^*_g + d_{1gi}y_{gi0}a_g + \tilde{X}_{gi}\beta_{gi} + d_{1gi}\tilde{\omega}_{gi} + d_{2gi}\bar{x}'_{gi}c^*_g + \varepsilon_{gi}, \quad i = 1, ..., N_g. \tag{12.51}$$

where $\tilde{y}_{gi} = [y_{gi0}, y_{gi1}, ..., y_{giT}]'$, $\varepsilon_{gi} = [\varepsilon_{gi0}, \varepsilon_{gi1}, ..., \varepsilon_{giT}]'$, $\tilde{X}_{gi} = diag(\tilde{x}'_{git})$ with $\tilde{x}_{git} = [0, x_{git}]$, $\bar{x}_{gi} = \sum_t^T T^{-1}(x_{git})$, $d_{1gi} = [d_{1git}]$, $d_{1git} = 1$ if t time period different to 0 and 0 otherwise, $d_{2gi} = [d_{2git}]$, $d_{2git} = 1$ if t time period equal to 0 and 0 otherwise.

$$B_g = \begin{bmatrix} 1 & 0 & 0 & ... & 0 & 0 \\ -\rho_g & 1 & 0 & 0 & ... & 0 \\ 0 & -\rho_g & 1 & 0 & ... & 0 \\ ... & & & & & \\ . & & & & & \\ . & & & & & \\ 0 & 0 & 0 & ... & -\rho_g & 1 \end{bmatrix}.$$

Pre-multiplying model 12.51 by $\Gamma_g = B_g^{-1}$ we obtain:

$$\tilde{y}_{gi} = \Gamma_g d_{1gi}\bar{x}'_{gi}b^*_g + \Gamma_g d_{1gi}y_{gi0}a_g + \Gamma_g d_{1gi}\tilde{X}_{gi}\beta_{gi} + \Gamma_g d_{1gi}\tilde{\omega}_{gi} + \Gamma_g d_{2gi}\bar{x}'_{gi}c^*_g + \Gamma_g\varepsilon_{gi}, \quad i = 1, ..., N_g. \tag{12.52}$$

with $\Gamma_g = \begin{bmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ \rho_g & 1 & 0 & 0 & ... & 0 \\ \rho_g^2 & \rho_g & 1 & 0 & ... & 0 \\ \rho_g^3 & \rho_g^2 & \rho_g & 1 & ... & 0 \\ ... & & & & & \\ \rho_g^{T-1} & \rho_g^{T-2} & \rho_g^{T-3} & ... & \rho_g & 1 \end{bmatrix}.$

The expression of B_g^{-1} is presented by Moreira (2009).

Now, we can set up the discrepancy function between the variance-covariance of the transformed augmented model 12.52 and its sample estimator:

$$\ell_g = log|\Sigma_g| + tr[S_g\Sigma_g^{-1}], \tag{12.53}$$

where

$$S_g = \frac{1}{N_g}\sum_{gi}(\varkappa_{gi}\varkappa'_{gi}), \tag{12.54}$$

with $\varkappa_{gi} = \tilde{y}_{gi} - \Gamma_g d_{1gi}\bar{x}'_{gi}b^*_g - \Gamma_g d_{1gi}y_{gi0}a_g - \Gamma_g d_{1gi}\tilde{X}_{gi}\beta_{gi} - \Gamma_g d_{1gi}\tilde{\omega}_{gi} - \Gamma_g d_{2gi}\bar{x}'_{gi}c^*_g$, $d_{2,gi}$ equal to 1 if t equal to 0 and 0 otherwise, $d_{1,gi}$ equal to 1 if t equal to 1 and

0 otherwise, $\Sigma_g = E[S_g]$. In addition, notice that $\varkappa_{gi} = \Gamma_g \iota_{T_{i_g}}(\tilde{\omega}_{gi}) + \Gamma_g(u_{gi})$ with $u_{gi} = d_{1gi}\tilde{x}_{gi}\lambda_{gi} + \varepsilon_{gi}$.

We estimate the cluster specific parameters of interest (ρ_g, β_g, $\triangle_g$, $\sigma^2_{\omega,g}$, and $\sigma^2_{\varepsilon,g}$) by maximizing the discrepancy function ℓ_g multiplied by $-N_g/2$. For given Σ_g, the estimator of the coefficients of interest is the GLS estimator per cluster:

$$\theta_g = (\tilde{\tilde{X}}'_g \Sigma_g \tilde{\tilde{X}}_g)^{-1}(\tilde{\tilde{X}}'_g \Sigma_g \tilde{y}_g), \tag{12.55}$$

where $\theta_g = [\rho_g, \beta_g, a_g, b^*_g, c^*_g]$, $\tilde{\tilde{X}}_g = [\tilde{y}_{g,t-1}, d_{1g}\tilde{X}_g, d_{1g}y_{0,g}, d_{1g}\bar{x}_g, d_{2g}\bar{x}_g]$.

For estimation of the variance-covariance matrix, we need to modify the estimator proposed in Subsection 4.1. The modified procedure is presented below.

Covariance Estimation In order to make GLS feasible, we propose the following estimation method of the variance-covariance components of $\triangle_{\lambda_g}$, $\sigma^2_{\tilde{\omega}_g}$, $\sigma^2_{\varepsilon_{0,g}}$, and $\sigma^2_{\varepsilon_g}$. More specifically, Σ_g is equal to the following block diagonal matrix:

$$\Sigma_g = \begin{pmatrix} \sigma^2_{\varepsilon_0,g} & 0 \\ 0 & \tilde{\Omega}_g \end{pmatrix} \tag{12.56}$$

where $\tilde{\Omega}_g = \sigma^2_{\tilde{\omega}}g(I_{N_g} \otimes \iota_T) + diag(X_g)(I_{N_gT} \otimes \Delta_{\lambda_g})diag(X_g) + \sigma^2_{\varepsilon_g} I_{N_g}$ if $T_g = T$. if $T_g \neq T$, one just need to set up the adequate design matrix for allowing unbalancedness in the time dimension.

First, we derive the linear decomposition of the variance-covariance matrix for each cluster:

$$\Sigma_g = \sum_{k=1}^{K}\sum_{k'=1}^{K} \sigma_{\lambda_g,kk'} H_{g,kk',\lambda_g} + \sigma^2_{\epsilon_g} I_{n_g} + \sigma^2_{\tilde{\omega}}g(I_{N_g} \otimes \iota_T). \tag{12.57}$$

with the design matrices equal to:

$$H_{g,kk',\lambda_g} = \tilde{X}_{g,k}\tilde{X}'_{g,k'},$$

where $\tilde{X}_{g,k} = diag(x_{git,k})$.

The estimation procedure is different to the one presented in subsection 4.1 because the model presents cluster-individual fixed effects.

Step 1: we obtain $\tilde{\tilde{y}}_g = y_g - y_{g,-1}\hat{\rho}_{GMM} - X_g\hat{\beta}_{g,GMM}$ where GMM stands for first-difference GMM estimation.

Step 2: we regress $\tilde{\tilde{y}}_g$ on $\iota_T \otimes \bar{x}_{gi}$, and $\iota_T y_{gi0}$.

Step 3: we obtain the residuals of regression of step 2, and call them r_g.

Step 4: we regress y_{gi0} on $\bar{x}_{gi}$, and we obtain the residuals calling them r_{g0}.

Step 5: we use the residuals $\tilde{r}_g = [r_{g0}, r_g]$.

Then, it can be shown that:

$$E[\tilde{r}_g\tilde{r}'_g] = \Sigma_g. \tag{12.58}$$

Replacing expression (4.10) into equation (12.58) and applying the vec operator, we obtain:

$$vec(E[\tilde{r}_g\tilde{r}_g']) = \sum_{k=1}^{K}\sum_{k'=1}^{K} \sigma_{\lambda_g,kk'} vec(H_{g,kk',\lambda_g}) + \sigma^2_{\epsilon_g} vec(I_{N_gT}) + \sigma^2_{\tilde{\omega}_g} vec(I_{N_g} \otimes \iota_T). \tag{12.59}$$

Now, we can rewrite the previous expression in matrix form:

$$vec(E[\tilde{r}_g\tilde{r}_g']) = \tilde{B}_{\lambda_g} vec(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(I_{N_gT}) + \sigma^2_{\tilde{\omega}_g} vec(I_{N_g} \otimes \iota_T). \tag{12.60}$$

In order to avoid double estimation of the covariances in the variance-covariance matrix, we use the identity $vec(A) = Dvech(A)$ where is A is square symmetric matrix and we re-express the previous equation as:

$$vec(E[\tilde{r}_g\tilde{r}_g']) = \tilde{B}_{\lambda_g} vech(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vech(I_{N_gT}) + \sigma^2_{\tilde{\omega}_g} vech(I_{N_g} \otimes \iota_T). \tag{12.61}$$

The expectation of the outer product of the residuals is replaced by the point estimator of the OLS residuals for each cluster and we add the error $\nu_{2,g}$ that captures the sampling error.

$$vec(\tilde{r}_g\tilde{r}_g') = \tilde{B}_{\lambda_g} vech(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(I_{N_gT}) + \sigma^2_{\tilde{\omega}_g} vech(I_{N_g} \otimes \iota_T) + \nu_{2,g}. \tag{12.62}$$

Finally, notice that 12.62 is a simple linear model that can be rewritten as:

$$\tilde{R}_g = \tilde{C}_g\tilde{\eta}_g + \nu_{2,g},$$

where:

$$\tilde{R}_g = vec(\tilde{r}_g\tilde{r}_g'),$$

$$\tilde{C}_g = [\tilde{B}_{\lambda_g}D \quad vec(I_{N_gT}) \quad vech(I_{N_g} \otimes \iota_T)],$$

$$B_{\lambda_g} = [vec(H_{g,11,\lambda_g}) \quad vec(H_{g,12,\lambda_g}) \quad ... \quad vec(H_{g,KK,\lambda_g})],$$

$$\eta_g = [vech(\triangle_{\lambda_g})' \quad \sigma^2_{\epsilon_g}]'.$$

Now, the estimators of the elements of variance-covariance are obtained by minimizing the following penalized loss function:

$$L(\eta_g) = (\tilde{R}_g - \tilde{C}_g\eta_g)'(\tilde{R}_g - \tilde{C}_g\eta_g) + \tau \parallel \tilde{\eta}_g \parallel_2^2,$$

with $\tau \in [0, 2min(\zeta_{gl})]$ where ζ_{gl} is the eigenvalue l of the matrix $\tilde{C}_g'\tilde{C}_g$.

Step 6: Iterate until convergence.

Remark 12.3. If we assume that λ_{git} is not present, we can estimate the cluster-specific parameters using different estimation procedures. If we keep the model in levels and we consider the cluster heterogeneity, we can use the IFGLS estimator of Phillips (2010), or the Within debiased estimator proposed by Breitung et al. (2022). If we first-difference the model within clusters, we can use the augmented estimator proposed by Chudik and Pesaran (2022) to estimate the cluster-specific parameters.

13. HOW ABOUT CROSS-SECTIONAL DEPENDENCE?

13.1. A model including common factors

The models 1.1 and 12.36 do not consider cross-sectional correlation even though cross-sectional dependence is a common problem in panel data.

Cross-sectional dependence is caused by spatial dependence or common shocks (Bai and Li (2021)) and it can be modeled either using spatial or factor models or a combination of both.

In this section, we extend model 1.1 in order to allow for cross-sectional dependence using a factor model. For this purpose, we include a cluster-time-specific fixed effect since it represents a cluster common factor. This is possible because the cluster-time specific effect τ_{gt} can be rewritten as $\sum_i^m s_i^{(g)} f_{gt}$ with $s_i^{(g)}$ equal to 1 and 0 (Bonhomme and Manresa (2015), Kapetanios et al. (2017), Bai and Li (2021)). Additionally, we include time-specific effects that capture common global factors across clusters.

The extended model 3.2 includes cluster-time additive effects as well as time-fixed effects as common factors for individual i in cluster g:

$$y_{git} = \alpha_g + \gamma_t + \tau_{gt} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}, \tag{13.63}$$

In this setting, Assumption 3.12 is relaxed to allow for regressors that present common factors.

Assumption 13.1. *x_{git} are generated from:*

$$x_{git} = \mu_g + \gamma_t + \tau_{gt} + \rho_x x_{git-l} + \omega_{git}, \qquad |\rho_x| < 1.$$

13.2. Identification and Estimation

The Mean Cluster estimator presented in section 4 estimates consistently the parameters of interest of model 13.63, which includes time and cluster-time dummies, by exploiting the different moment conditions derived in this subsection.

We obtain moment conditions using the deviations with respect to cluster-time specific averages:

$$\begin{aligned} y_{git} - y_{g.t} = \rho_g(y_{git-1} - y_{g.t-1}) + (x_{git} - x_{g.t})'\beta_g \\ + x'_{git}\lambda_{git} - x'_{g.t}\lambda_{g.t} + \varepsilon_{git} - \varepsilon_{g.t}. \end{aligned} \tag{13.64}$$

The cluster-time specific averages are equal to:

$$\frac{\sum_i y_{git}}{N_g} = \alpha_g + \gamma_t + \tau_{gt} + \rho_g \frac{\sum_i y_{git-1}}{N_g} + \frac{\sum_i x_{git}}{N_g}\beta_g + \frac{\sum_i x'_{git}\lambda_{git}}{N_g} + \frac{\sum_i \varepsilon_{git}}{N_g}. \tag{13.65}$$

We can just rename the transformed variables as:

$$\tilde{y}_{git} = \rho_g \tilde{y}_{git-1} + \tilde{x}'_{git}\beta_g + \widetilde{x'_{git}\lambda_{git}} + \tilde{\varepsilon}_{git}. \tag{13.66}$$

Thus, after this transformation we obtain the following moment conditions:

$$E(\tilde{u}_{git}\tilde{x}_{gis}) = 0, \quad s = 1, 2, ..., T, \quad i = 1, 2, ..., N_g, \quad g = 1, 2, ..., m, \tag{13.67}$$

$$E(\tilde{u}_{git}\tilde{y}_{git-1}) = 0, \quad t = 1, 2, ..., T, \quad i = 1, 2, ..., N_g, \quad g = 1, 2, ..., m. \tag{13.68}$$

In addition, we need to add the full-rank condition for the transformed regressors.

Assumption 13.2. *OLS: The matrix $E(\tilde{z}_{git}\tilde{z}'_{git})$ is full rank.*

GLS: $E[\tilde{u}_g\tilde{u}'_g]$ is positive definite and $E(\tilde{Z}'_g E[\tilde{u}_g\tilde{u}'_g]^{-1}\tilde{Z}_g)$ is nonsingular.

14. LONG TIME DIMENSION

Until now, we have focused on a dynamic panel data model with clustering and short-time dimension. As mentioned, the problems in this setting are incidental parameters and initial conditions dependency. When the time dimension is long, one does not run on the problem of initial conditions dependency but the issues of non-stationarity and incidental parameter bias are still important.

More specifically, if the time dimension is long the assumption that the initial conditions are generated from the stationary distribution is no longer needed (Assumption 3.11). The reason is that the influence of the initial conditions becomes negligible as $T \to \infty$.

In addition, the assumptions of stationary regressors and stationary dependent variables are necessary for the consistency and asymptotical normality of the Mean Cluster estimator. The reason is that the stationarity of the regressors guarantees that the error term of the model is integrated of order 0. In the case of non-stationary regressors, we have two options: 1. transform the regressors to obtain stationarity or 2. estimate the model in levels if there is co-integration between the dependent variable and the regressors after including the lag of the dependent variable (Hamilton (1994)). In the last case, the assumption of cluster-additive effects is crucial to obtain asymptotically normal estimates within cluster (Choi (2015)). While in the presence of cluster-individual additive effects, it is necessary to use the fully-modified OLS estimator proposed by Phillips and Moon (1999) cluster per cluster or one can use OLS estimation with the Mundlak approach. If the dependent variable and the regressors are not cointegrated and the model presents cluster-specific additive effects, it is unclear if cluster OLS estimation and the Mean-Cluster estimator are consistent. The reason is that Phillips and Moon (1999) show that pooled OLS is a consistent estimator of the long-run average regression coefficient if the regressors are nonstationary and there is no cointegration for a model without intercept and lagged dependent variable. Thus, further research is needed to verify the consistency of the Mean-Cluster estimator when there is no co-integration and the model presents cluster additive specific effects. However, the MC-OLS estimator is consistent if the model presents additive cluster-individual specific effects, and there is no cointegration if we use the Mundlak approach (Phillips and Moon (1999)). Finally, in order to test for co-integration one can extend the test proposed by Im et al. (2003) such that the model presents cluster-specific parameters instead of individual-specific parameters. Concluding that there is co-integration would entail that $u_{git} = x'_{git}\lambda_{git} + \varepsilon_{git}$ is stationary, meaning that λ_{git} could be considered as a random co-integrating vector. A study of a co-integration test and the properties of the Mean-Cluster estimator when there is no co-integration is out of the scope of this paper and both issues are left for further research.

On the other hand, the problem of incidental parameter bias requires careful analysis. First, the problem of incidental parameter bias in model 3.2 is not present. The intuitive explanation is that we have increasing observations to estimate cluster-specific parameters. But if we allow for cluster-individual specific effects as in model 12.36, we need to be more careful. In this setting, the estimated cluster-specific parameters using the within estimator are consistent and asymptotically normal if $lim\frac{N_g}{T_{ig}} = 0$ and the regressors are not stationary (Phillips and Moon (1999)). If the regressors are stationary, we must debias the within estimator per cluster (Hahn and Newey (2004)). A workaround to avoid the condition $lim\frac{N_g}{T_{ig}} = 0$ or debiasing is to use the Mundlak approach and project the cluster-individual specific effects into the column space of the regressors. Finally, model 13.63 suffers the problem of incidental parameter bias due to the presence of cluster-individual specific effects and cluster-time specific effects. But the transformation proposed in subsection 13.2 eliminates the incidental parameter problem.

Finally, the Mean-Cluster estimator and the Mean-Group estimator are consistent estimators of the mean coefficients of model 3.2 when the time dimension (Subsection 8.3).

15. UNKNOWN CLUSTERING?

In this paper, we focused on known clustering because clustering with short T is not feasible. The reason is that we only have T observations to determine the membership of individual i into one of m possible groups (Bonhomme and Manresa (2015), Sarafidis and Wansbeek (2021)). To be more specific, Bonhomme and Manresa (2015) show that the Group Fixed Effects estimator converges to a pseudo value that might not be equal to the true parameter when the time dimension is short.

On the contrary, there are several available methodologies when the time dimension is long. One can find the Grouped Fixed Effects (GFE) estimator for a model with homogeneous slopes and cluster-time additive effects proposed by Bonhomme and Manresa (2015) that is consistent when the number of individuals and the number of time periods grow to infinity. In addition, they show that their estimator is suited for a time dimension as short as 7. In addition, Bonhomme and Manresa (2015) explain that group misclassification produces higher finite sample dispersion of the estimator. Finally, Bonhomme and Manresa (2015) explain that the GFE estimator for a model with cluster-specific slopes is consistent only when the time dimension grows to infinity.

Similarly, Su et al. (2016) developed a classifier-Lasso that allows the estimation of unknown group-specific parameters when group membership is unknown. This estimator is consistent when the number of individuals and the number of time observations grow to infinity.

16. MEASUREMENT ERROR

In this Section, we focus on a model with measurement error in the dependent variable. We assume that the observed dependent variable is given by:

$$y^*_{git} = y_{git} + \epsilon_{git}, \tag{16.69}$$

with y_{git} the true process, and ϵ_{git} the measurement error with $E[\epsilon_{git}] = 0$, and $E[\epsilon^2_{git}] = \sigma^2_\epsilon$.

Since the lag of the observed dependent variable is included in the right-hand side of the model, we end up with an endogeneity issue. Replacing $y_{git} = y^*_{git} - \epsilon_{git}$ into model 3.2, we obtain:

$$y^*_{git} - \epsilon_{git} = \rho_g y^*_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git} + \alpha_{gi} - \epsilon_{git-1}. \tag{16.70}$$

In model 16.70, it is easy to see that the lag of the observed dependent variable is correlated with the new composite error term $\varepsilon_{git} - \epsilon_{git-1}$.

If we first-difference the model, we obtain:

$$\Delta y^*_{git} - \Delta\epsilon_{git} = \rho_g \Delta y^*_{git-1} + \Delta x'_{git}\beta_{git} + \Delta\varepsilon_{git} - \rho_g\Delta\epsilon_{git-1}. \tag{16.71}$$

In the first-differenced model 16.74, the first-differenced observed lagged dependent variable is correlated with the error term $\Delta\varepsilon_{git} - \Delta\epsilon_{git-1}$. If $T = 3$, we propose to use the following moment conditions for identification of the parameters of interest:

$$E[x_{git-1}\Delta\epsilon_{git-1}] = 0 \tag{16.72}$$

$$E[x_{git-2}\Delta\varepsilon_{git}] = 0 \tag{16.73}$$

Another option is to use the model in levels conditional in the initial observed values; in this case, we have the following model:

$$y^*_{git} - \Delta\epsilon_{git} = \rho_g y^*_{git-1} + x'_{git}\beta_{git} + \theta_g y_{gi0} + \varepsilon_{git} + \alpha_{gi} - \rho_g\epsilon_{git-1}. \tag{16.74}$$

This model can be estimated using an FGLS-IV approach exploiting the instruments. If $T = 3$, we can only use as instruments x_{git-1} as an instrument for y_{git-1} and x_{gi0} as an instrument of y_{gi0}.

17. MONTE CARLO EXPERIMENT: STRATIFIED SAMPLING

In this section, we present a Monte Carlo simulation experiment to test the proposed estimators for the baseline model and the extensions of the baseline model under stratified sampling.

For this purpose, we generate 100 datasets from five different data-generating processes called DGP 1, DGP 2, DGP 3, and DGP 4. We use DGP 1 to test the proposed Mean Cluster estimator under the assumption of clustered unobserved heterogeneity, DGP 2 to test the Bayesian estimator, DGP 3 to test the Mean Cluster estimator in the presence of time effects, and DGP 4 to test the Mean FGLS estimator proposed for a model with inclusion of additive cluster-individual effects.

In the following subsections, we describe the different designs in more detail as well as the results.

17.1. The design

17.1.1. DGP 1 In order to test the Mean Cluster estimator proposed for a model with cluster unobserved heterogeneity and mixed coefficients (model 1.1), we conduct a simulation experiment using a data-generating process that is similar to the DGP used by Arellano and Bond (1991). We use a modification of the DGP proposed by Arellano and Bond (1991) to illustrate that the first-differenced GMM estimator breaks down in the presence of multiplicative unobserved cluster heterogeneity.

The main differences with the DGP of Arellano and Bond (1991) are: 1. inclusion of cluster additive effects instead of individual-specific effects that are correlated with the regressors, 2. inclusion of multiplicative cluster-individual-time specific effects, 3. the variance and variance-covariance are cluster-specific and they are generated from Gamma and Wishart distributions.

More specifically, we generate 100 samples from the following model for individual i in cluster g at period t:

$$y_{git} = \alpha_{1,g} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git},$$

with $\rho_g = \bar{\rho} + \alpha_{2,g}$, $\bar{\rho}$, $\beta_{git} = \bar{\beta} + \alpha_{3,g} + \lambda_{git}$ and $\bar{\beta} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

The number of clusters is equal to 4, the number of individuals within cluster is equal to 100, and the number of time observations is equal to 3.

The cluster additive effects $\alpha_{1,g}$ are generated from a normal distribution centered at 0 with heteroskedastic variance across clusters ($\sigma^2_{\alpha,1g} \in \{1.01, 1.01, 0.9, 0.9\}$).

The cluster effects ($\alpha_{2,g}$) added to the persistence parameter ($\bar{\rho}$) are centered at 0, and equal to $\alpha_{2,g} \in \{-0.5, -0.5, 0.5, 0.5\}$.

The cluster effects ($\alpha_{3,g}$) added to the mean coefficient vector ($\bar{\beta}$) are centered at 0, and equal to $\alpha_{3,g} \in \{-0.5, -0.5, 0.5, 0.5\}$.

The cluster-individual-time specific effects (λ_{git}) added to the mean coefficient vector ($\bar{\beta}$) are generated from a multivariate normal distribution centered at 0 with heteroskedastic variance-covariance matrix across clusters ($\Delta_{\lambda,1} = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$, $\Delta_{\lambda,2} = \begin{pmatrix} 0.11 & 0.05 \\ 0.05 & 0.11 \end{pmatrix})$, $\Delta_{\lambda,3} = \begin{pmatrix} 0.12 & 0.05 \\ 0.05 & 0.12 \end{pmatrix})$, $\Delta_{\lambda,4} = \begin{pmatrix} 0.13 & 0.05 \\ 0.05 & 0.13 \end{pmatrix})$.

The disturbance term (ε_{git}) is generated from a normal distribution centered at 0 with a cluster heteroskedastic variance ($\sigma^2_{\varepsilon g} \in \{0.9, 0.9, 1.01, 1.01\}$).

The regressors x_{git} follow stationary autoregressive processes similar to the process used by Arellano and Bond (1991). The key difference is that we allow for correlation with the cluster effects:

$$x_{git} = \alpha_{1,g} + \alpha_{3,g} + \phi x_{git-1} + \omega_{git},$$

with ϕ is equal to 0.8.

The disturbance term of the regressors (ω_{git}) equation is sampled from the normal distribution centered at 0 with variance that is cluster specific ($\sigma^2_{\omega_g} \in \{0.9, 0.9, 1.01, 1.01\}$).

17.1.2. DGP 2 In order to test the estimator proposed in subsection 12.2.2, we conduct a simulation experiment using a data-generating process similar to the DGP used by Arellano and Bond (1991).

The main differences with the DGP of Arellano and Bond (1991) are: 1. inclusion of correlated cluster-individual effects instead of individual-specific effects, 2. inclusion of multiplicative cluster-individual-time specific effects, 3. the variance and variance-covariance are cluster-specific and they are generated from Gamma and Wishart distributions.

More specifically, we generate 100 samples from the following model for individual i in cluster g at period t:

$$y_{git} = \alpha_{1,gi} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git},$$

with $\rho_g = \bar{\rho} + \alpha_{2,g}$, $\bar{\rho}$, $\beta_{git} = \bar{\beta} + \alpha_{3,g} + \lambda_{git}$ and $\bar{\beta} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$.

The number of clusters is equal to 4, the number of individuals within cluster is equal to 100, and the number of time observations is equal to 3.
The cluster-individual additive effects $\alpha_{1,gi}$ are generated from a normal distribution centered at 0 with variance equal to 1.
The cluster effects ($\alpha_{2,g}$) added to the persistence parameter ($\bar{\rho}$) are centered at 0, and equal to $\alpha_{2,g} \in \{-0.025, -0.025, 0.025, 0.025\}$.
The cluster effects ($\alpha_{3,g}$) added to the mean coefficient vector ($\bar{\beta}$) are centered at 0, and equal to $\alpha_{3,g} \in \{-0.05, -0.05, 0.05, 0.05\}$.
The cluster-individual-time specific effects (λ_{git}) added to the mean coefficient vector ($\bar{\beta}$) are generated from a multivariate normal distribution centered at 0 with with variance-covariance matrix equal across clusters ($\Delta_{\lambda,1} = \begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$.
The disturbance term (ε_{git}) is generated from a normal distribution centered at 0 with a cluster heteroskedastic variance ($\sigma^2_{\varepsilon g} \in \{0.9, 0.9, 1.01, 1.01\}$).

The regressors x_{git} follow stationary autoregressive processes similar to the process used by Arellano and Bond (1991). The key difference is that we allow for correlation with the cluster effects:

$$x_{git} = \alpha_{1,gi} + \alpha_{3,g} + \phi x_{git-1} + \omega_{git},$$

with ϕ is equal to 0.8.
The disturbance term of the regressors (ω_{git}) equation is sampled from the a normal distribution centered at 0 with variance equal to 0.9.

17.1.3. DGP 3 In order to test the estimator proposed in Section 16, we conduct a simulation experiment using a data generating processes that is similar to the DGP 3 described above. The only difference is that the observed dependent variable is given by:

$$y^{obs}_{git} = y_{git} + \epsilon_{git}, \tag{17.75}$$

where ϵ_{git} represents the measurement error generated from a standard normal distribution.

17.2. The Results

17.2.1. DGP 1 In Table 1, we present the bias and RMSE of the estimated mean parameters of interest for different values of the persistence parameter. The estimates are obtained for 100 simulations for a sample with 4 clusters, 100 individuals per group, and a time dimension equal to 3.

The results show that the proposed Mean Cluster FGLS estimators have lower bias and RMSE than the first-differenced GMM estimators.

17.2.2. DGP 2 In Table 2, we present the bias and RMSE of the estimated mean parameters of interest for different values of the persistence parameter. The estimates are obtained for 100 simulations for a sample with 4 clusters, 100 individuals per group and a time dimension equal to 3. The results show that the proposed Mean Cluster FGLS estimators have lower Bias and RMSE than the first-differenced GMM estimators and the system GMM estimators.

Table 1: DGP 1

	Mean	**Bias**	**RMSE**	**Mean**	**Bias**	**RMSE**	**Mean**	**Bias**	**RMSE**
	$\rho = 0.1$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.0931	-0.0069	0.001	1.0055	0.0055	0.0057	1.0055	0.0055	0.0062
MC-OLSy0	0.0931	-0.0075	0.002	1.0066	0.0066	0.0056	1.0066	0.0066	0.0064
MC-FGLS	0.0908	-0.0092	0.0123	1.0314	0.0314	0.0782	1.0314	0.0314	0.0241
MC-FGLSy0	*0.0988*	*-0.0012*	*0.0103*	*1.0073*	*0.0073*	*0.0246*	*1.0073*	*0.0073*	*0.9594*
MC-JIFDGMM	0.5545	0.4545	82.4716	0.8048	-0.1952	15.9684	0.8048	-0.1952	8.6445
JIFDGMM	0.1727	0.0727	3.4936	0.9846	-0.1952	0.0729	0.9846	-0.1952	0.3024
MC-OIGMM	0.0995	-0.0005	0.008	1.0919	0.0919	0.822	1.0919	0.0919	0.8789
OIGMM	0.1139	0.0139	0.0116	1.2076	0.2076	1.6085	1.2076	0.2076	1.3103
MC-SYSGMM	0.0901	-0.0099	0.0034	1.0851	0.0851	0.0303	1.0851	0.0851	0.0328
SYSGMM	0.1135	0.0135	0.0047	1.0937	0.0937	0.0356	1.0937	0.0937	0.0426
FGLS-Hsiao	-0.0999	-0.1999	0.0434	1.0156	0.0156	0.0149	1.0156	0.0156	0.017
	$\rho = 0.5$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.4931	-0.0069	0.0005	1.0083	0.0083	0.0061	1.0083	0.0083	0.0064
MC-OLSy0	0.4931	-0.0101	0.0021	1.009	0.009	0.0059	1.009	0.009	0.0065
MC-FGLS	0.4941	-0.0059	0.0081	1.0197	0.0197	0.1547	1.0197	0.0197	0.1069
MC-FGLSy0	*0.4947*	*-0.0053*	*0.0261*	*1.0028*	*0.0028*	*0.0649*	*1.0028*	*0.0028*	*0.0924*
MC-JIFDGMM	1.4552	0.9552	144.1671	1.0147	0.0147	5.5668	1.0147	0.0147	5.7164
JIFDGMM	0.5373	0.0373	0.4727	0.9605	0.0147	0.0566	0.9605	0.0147	0.0507
MC-OIGMM	0.4861	-0.0139	0.0107	0.9989	-0.0011	0.7261	0.9989	-0.0011	0.841
OIGMM	0.4962	-0.0038	0.0193	1.0505	0.0505	1.4267	1.0505	0.0505	1.1413
MC-SYSGMM	0.4894	-0.0106	0.0025	1.0912	0.0912	0.0318	1.0912	0.0912	0.0359
SYSGMM	0.5257	0.0257	0.0043	1.0782	0.0782	0.0382	1.0782	0.0782	0.0438
FGLS-Hsiao	0.2641	-0.2359	0.0594	1.0283	0.0283	0.0156	1.0283	0.0283	0.0186
	$\rho = 0.9$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.8974	-0.0026	0.0002	1.0053	0.0053	0.0061	1.0053	0.0053	0.0064
MC-OLSy0	0.8974	-0.0062	0.0013	1.0064	0.0064	0.006	1.0064	0.0064	0.0065
MC-FGLS	0.9036	0.0036	0.0043	1.0246	0.0246	0.2824	1.0246	0.0246	0.1272
MC-FGLSy0	*0.9285*	*0.0285*	*0.0649*	*1.0095*	*0.0095*	*0.1402*	*1.0095*	*0.0095*	*0.0832*
MC-JIFDGMM	0.9536	0.0536	0.1542	0.9805	-0.0195	0.0363	0.9805	-0.0195	0.1542
JIFDGMM	0.9122	0.0122	0.0031	0.9698	-0.0195	0.032	0.9698	-0.0195	0.0386
MC-OIGMM	0.8783	-0.0217	0.0075	1.0368	0.0368	1.1681	1.0368	0.0368	1.0338
OIGMM	0.8996	-0.0004	0.0052	1.2316	0.2316	1.5646	1.2316	0.2316	1.428
MC-SYSGMM	0.9015	0.0015	0.0005	1.0766	0.0766	0.0316	1.0766	0.0766	0.0331
SYSGMM	0.9256	0.0256	0.0014	1.0576	0.0576	0.0403	1.0576	0.0576	0.0407

Note: MC-OLS: Mean-Cluster OLS estimator, MC-FGLS: Mean-Cluster FGLS estimator,
MC-OLSy0: Mean-Cluster OLS estimator conditioning on initial value,
MC-FGLSy0: Mean-Cluster FGLS estimator conditioning on initial value,
MC-JIFDGMM: Mean-Cluster just-identified fist-differenced GMM estimator,
JIFDGMM: Just-identified fist-differenced GMM estimator,
MC-OIFDGMM: Mean-Cluster over-identified fist-differenced GMM estimator,
OIFDGMM: Over-identified fist-differenced GMM estimator, MC-SYSGMM: System GMM estimator, SYSGMM: System GMM estimator.

17.2.3. DGP 3 In Table 3, we present the bias and RMSE of the estimated mean parameters of interest for different values of the persistence parameter. The estimates are obtained for 100 simulations for a sample with 4 clusters, 100 individuals per group and a time dimension equal to 3. The results show that the proposed Mean Cluster FGLS-IV estimator has lower bias and RMSE than the first-differenced estimators.

18. CONCLUSIONS

In this paper, we investigate the identification and estimation of dynamic heterogeneous linear models in the presence of cluster heterogeneity when cluster structure is known and panel data is unbalanced due to randomly missing data with a short or fixed time dimension.

In order to exploit the structure of the data, this article proposes two approaches depending on the growth of the number of clusters. When the number of clusters is fixed, we observe all the clusters and the number of individuals grows to infinity, it

Table 2: DGP 2

	Mean	**Bias**	**RMSE**	**Mean**	**Bias**	**RMSE**	**Mean**	**Bias**	**RMSE**
	$\rho = 0.1$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.0935	-0.0065	0.0036	1.0055	0.0055	0.0172	1.0025	0.0025	0.0161
MC-FGLS	*0.1030*	*0.0030*	*0.0036*	*1.0046*	*0.0046*	*0.0089*	1.0046	0.0046	0.0093
MC-JIFDGMM	0.5649	0.4649	85.1471	1.4842	0.4842	14.9840	0.6957	-0.3043	8.0302
JIFDGMM	-0.4678	-0.5678	21.8730	1.1253	0.1253	1.2624	1.0540	0.0540	0.7525
MC-OIFDGMM	0.5649	0.4649	85.1471	1.4842	0.4842	14.9840	0.6957	-0.3043	8.0302
OIFDGMM	0.0749	-0.0251	0.0168	0.9938	-0.0062	1.0178	0.9364	-0.0636	1.2441
MC-SYSGMM	0.0723	-0.0277	0.0074	1.1414	0.1414	0.0400	1.1130	0.1130	0.0407
SYSGMM	0.0949	-0.0051	0.0088	1.1264	0.1264	0.0429	1.0925	0.0925	0.0404
	$\rho = 0.5$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.4921	-0.0079	0.0039	1.0191	0.0191	0.0171	1.0171	0.0171	0.0162
MC-FGLS	*0.4966*	*-0.0034*	*0.0288*	*1.0761*	*0.0761*	*0.2015*	*1.0190*	*0.0190*	*0.0357*
MC-FDGMM	-0.1210	-0.6210	108.3416	1.1141	0.1141	0.5770	0.8036	-0.1964	1.8841
FDGMM	0.5566	0.0566	0.1178	0.9929	-0.0071	0.0319	0.9725	-0.0275	0.0373
MC-OIGMM	-0.1210	-0.6210	108.3416	1.1141	0.1141	0.5770	0.8036	-0.1964	1.8841
OIGMM	0.4702	-0.0298	0.0217	1.1090	0.1090	0.8645	1.0439	0.0439	0.8128
MC-SYSGMM	0.4773	-0.0227	0.0056	1.1533	0.1533	0.0470	1.1282	0.1282	0.0526
SYSGMM	0.5159	0.0159	0.0062	1.1095	0.1095	0.0397	1.0724	0.0724	0.0503
	$\rho = 0.9$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.8976	-0.0024	0.0019	1.0264	0.0264	0.0167	1.0247	0.0247	0.0161
MC-FGLS	*0.8975*	*-0.0025*	*0.0009*	*1.0297*	*0.0297*	*0.0090*	*1.0233*	*0.0233*	*0.0082*
MC-JIFDGMM	0.9020	0.0020	0.0033	0.9972	-0.0028	0.0319	0.9825	-0.0175	0.0337
JIFDGMM	0.9088	0.0088	0.0028	0.9883	-0.0117	0.0306	0.9736	-0.0264	0.0359
MC-OIGMM	0.9020	0.0020	0.0033	0.9972	-0.0028	0.0319	0.9825	-0.0175	0.0337
OIGMM	0.8809	-0.0191	0.0074	1.4201	0.4201	1.6271	1.3970	0.3970	1.7727
MC-SYSGMM	0.9012	0.0012	0.0010	1.1236	0.1236	0.0381	1.0906	0.0906	0.0422

Note: MC-OLS: Mean-Cluster OLS estimator conditioning on the initial value,
MC-FGLS: Mean-Cluster FGLS estimator conditioning on the initial value,,
MC-JIFDGMM: Mean-Cluster just-identified fist-differenced GMM estimator,
JIFDGMM: Just-identified fist-differenced GMM estimator,
MC-OIFDGMM: Mean-Cluster over-identified fist-differenced GMM estimator,
OIFDGMM: Over-identified fist-differenced GMM estimator, MC-SYSGMM: System GMM estimator, SYSGMM: System GMM estimator.

Table 3: DGP 3

	Mean	**Bias**	**RMSE**	**Mean**	**Bias**	**RMSE**	**Mean**	**Bias**	**RMSE**
	$\rho = 0.1$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.0908	-0.0092	0.0006	1.0093	0.0123	0.0027	1.0093	0.0123	0.0028
MC-FGLSIV	*0.0914*	*-0.0086*	*0.0033*	*1.0051*	*0.0069*	*0.0039*	*1.0051*	*0.0069*	*0.0038*
MC-FDGMM	-0.0126	-0.1126	0.2135	1.015	0.015	0.0092	1.015	0.015	0.0114
	$\rho = 0.5$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.443	-0.057	0.0038	1.0101	0.0128	0.0027	1.0101	0.0128	0.0029
MC-FGLSIV	*0.4909*	*-0.0091*	*0.003*	*1.0074*	*0.0094*	*0.0043*	*1.0074*	*0.0094*	*0.0042*
MC-FDGMM	0.5138	0.0138	0.0109	1.0185	0.0185	0.0069	1.0185	0.0185	0.0098
	$\rho = 0.9$			$\beta_1 = 1$			$\beta_2 = 1$		
MC-OLS	0.8647	-0.0353	0.0015	1.0373	0.0401	0.0043	1.0373	0.0401	0.0043
MC-FGLSIV	*0.8938*	*-0.0062*	*0.0011*	*1.012*	*0.0144*	*0.0057*	*1.012*	*0.0144*	*0.0056*
MC-FDGMM	0.9006	0.0006	0.0005	1.0188	0.0188	0.0076	1.0188	0.0188	0.0103

Note: MC-OLS: Mean-Cluster OLS estimator, MC-FGLSIV: Mean-Cluster FGLS-IV estimator,
MC-FDGMM: Mean-Cluster fist-differenced GMM estimator

is possible to estimate the mean slope coefficients and persistence parameter using a Mean Cluster estimator that is an extension of the Mean Group estimator introduced by Pesaran and Smith (1995). When the square root of the number of clusters is growing at a lower rate than the growth of the number of individuals within a cluster, the Mean Cluster estimators estimate consistently the mean parameters.

As an extension of the baseline model, we consider a model with cluster-individual additive effects. In this setting, we suggest a hierarchical Bayesian estimation with a prior for the unknown initial conditions. In addition, we propose to condition on the initial values in order to avoid making assumptions about the data-generating processes of the initial conditions and the regressors.

A second extension is a model that allows for cross-sectional dependence by including a common factor for the whole population and a cluster-specific common factor. In this setting, the Mean Cluster OLS estimator using the time-demeaned regressors outperforms pooled OLS.

We can conclude from the simulation experiment, that the Mean Cluster estimators have lower Relative Bias and RMSE than the MG estimator and OLS estimator. This shows that one can exploit the underlying clustering in the data to estimate the mean coefficients and the cluster-specific parameters of heterogeneous linear dynamic panel data models.

Finally, we show that the first-difference GMM estimator is inconsistent when there is multiplicative cluster heterogeneity. In fact, the first-difference GMM estimator is equal to the weighted average of the cluster-specific marginal effects. A similar conclusion can be drawn if the marginal effects are individual-specific. In addition, we show that the Mean Group estimator is equal to the Mean Cluster estimator when the time dimension is long and the data is obtained by means of stratified sampling.

19. ANNEX

19.1. Proofs of theorems 6.1,6.2,6.3

19.1.1. Proof Theorem 6.1

Proof. The cluster-specific GLS estimator is given by:

$$\hat{\theta}_{g,GLS} = (Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}y_g).$$

We can re-write it as follows:

$$\hat{\theta}_{g,GLS} = \theta_g + (Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g),$$

with:

$$w_g = diag(X_g)\lambda_g + \epsilon_g.$$

Applying the plim operator and using Slutksy's Theorem we obtain:

$$\underset{\substack{N_g\to\infty \\ T_{ig}\,fixed}}{plim}\hat{\theta}_{g,GLS} = \theta_g + (\underset{\substack{N_g\to\infty \\ T_{ig}\,fixed}}{plim}\frac{1}{n_g}Z_g'\Omega_g^{-1}Z_g)^{-1}(\underset{\substack{N_g\to\infty \\ T_{ig}\,fixed}}{plim}\frac{1}{n_g}Z_g'\Omega_g^{-1}w_g),$$

Now, $\underset{\substack{N_g\to\infty \\ T_{ig}\,fixed}}{plim} n_g^{-1}Z_g'\Omega_g^{-1}Z_g = Q_g$ by Assumption 5.1, and $\underset{\substack{N_g\to\infty \\ T_{ig}\,fixed}}{plim}\frac{1}{n_g}Z_g'\Omega_g^{-1}w_g = 0$ by Assumptions 3.8 and 3.9. The last conclusion is obtained as follows:

$\underset{\substack{N_g\to\infty \\ T_{i_g} fixed}}{plim} \frac{1}{n_g} Z_g' \Omega_g^{-1} w_g = \underset{\substack{N_g\to\infty \\ T_{i_g} fixed}}{plim} \sum_{i_g} \sum_{t_{i_g}} z_{i_g t_{i_g}} \omega_{i_g t_{i_g}} \sigma_{i_g t_{i_g}}$ because Ω_g^{-1} is a diagonal matrix with $\sigma_{i_g t_{i_g}}$ in each element of the diagonal. Then,

$$\underset{\substack{N_g\to\infty \\ T_{i_g} fixed}}{plim} \frac{1}{n_g} \sum_{i_g} \sum_{t_{i_g}} z_{i_g t_{i_g}} \omega_{i_g t_{i_g}} \sigma_{i_g t_{i_g}} = \frac{1}{T_{i_g}} \sum_{t_{i_g}} E_g[z_{i_g t_{i_g}} \omega_{i_g t_{i_g}} \sigma_{i_g t_{i_g}}],$$

where $E_g[z_{i_g t_{i_g}} \omega_{i_g t_{i_g}} \sigma_{i_g t_{i_g}}]$ represents the cross-sectional expectation. Now, under assumptions 3.8 and 3.9 $E_g[z_{i_g t_{i_g}} \omega_{i_g t_{i_g}} \sigma_{i_g t_{i_g}}] = 0$. Then, $\hat{\theta}_{g,GLS} - \theta_g = o_p(1)$.
In order to derive the asymptotic distribution of the cluster-specific parameter, I use the stabilizing factor equal to $\sqrt{n_g}$ such that:

$$\sqrt{n_g}(\hat{\theta}_{g,GLS} - \theta_g) = (\frac{1}{n_g} Z_g' \Omega_g^{-1} Z_g)^{-1} (\frac{1}{\sqrt{n_g}} Z_g' \Omega_g^{-1} w_g).$$

By Linderberg-Levy Central Limit Theorem $\frac{1}{\sqrt{n_g}} Z_g' \Omega_g^{-1} w_g \to N(0, Q_g)$.
Then,

$$\sqrt{n_g}(\hat{\theta}_{g,GLS} - \theta_g) \xrightarrow{d} N(0, Q_g^{-1}).$$

19.1.2. Proof Theorem 6.2

Proof. As derived in section 4.1, the variance-covariance components stacked up in the vector η_g are estimated by the penalized LS estimator as $\hat{\eta}_g = (C_g' C_g + \tau I)^{-1} (C_g' \hat{R}_g)$ with C_g a full rank matrix obtained following the procedure proposed there. Now, for $\tau = 0$:

$$\hat{\eta}_g = (C_g' C_g)^{-1} (C_g' \hat{R}_g)$$

That is equal to:

$$\hat{\eta}_g = \eta_g + (C_g' C_g)^{-1} (C_g' \nu_{git})$$

Now, applying the plim operator and using Slutsky's theorem:

$$plim \hat{\eta}_g = \eta_g + (plim \frac{1}{n_g} C_g' C_g)^{-1} (plim \frac{1}{n_g} C_g' \nu_g)$$

Now, $plim \frac{1}{n_g} C_g' C_g = D_g$ and $plim \frac{1}{n_g} C_g' \nu_g = 0$ because ν_g is an error capturing estimation error and it is orthogonal of C_g (Assumption 4.1). Thus, $\hat{\eta}_g = \eta_g$.
In order to derive the asymptotic distribution of the variance-covariance estimators, I use the stabilizing factor $\sqrt{n_g}$ such that:

$$\sqrt{n_g}(\hat{\eta}_g - \eta_g) = (\frac{1}{n_g} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} C_{i_g t_{i_g}} C_{i_g t_{i_g}}')^{-1} (\frac{1}{\sqrt{n_g}} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} C_{i_g t_{i_g}} \nu_{i_g t_{i_g}}).$$

Then, by Linderberg-Levy CLT we have that $\frac{1}{\sqrt{n_g}} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} C_{i_g t_{i_g}} \nu_{i_g t_{i_g}} \to N(0, \sigma_{\nu,g}^2 D_g)$.
Thus,

$$\sqrt{n_g}(\hat{\eta}_g - \eta_g) \xrightarrow{d} (0, \sigma_{\nu_g}^2 D_g^{-1}).$$

Finally, $\Omega_g = g(\triangle_{\lambda_g}, \sigma^2_{\varepsilon_g})$ and $g(.)$ is a continuous function because it is a linear decomposition. As a result, it is possible to use the Slutzky's theorem to such that:

$$\sqrt{n_g}(\hat{\Omega}_g - \Omega_g)\xrightarrow{d}N(0, var(\hat{\Omega}_g)).$$

19.1.3. Proof Theorem 6.3 [7]

Proof. We know that $\hat{\hat{\theta}} = \sum_g^G \hat{\pi}_g \hat{\theta}_g$. Since $\hat{\pi}_g = \pi_g + o_p(1)$ by Assumption 3.3, we can re-write the Mean Cluster estimator as follows:

$$\hat{\hat{\theta}} = \sum_g^m \pi_g \hat{\theta}_g + o_p(1).$$

In addition, we have that:

$$\hat{\hat{\theta}} = \sum_g^m \pi_g \theta_g + \sum_g^m \pi_g (Z_g' \Omega_g^{-1} Z_g)^{-1} (Z_g' \Omega_g^{-1} w_g) + o_p(1).$$

Then,

$$\hat{\hat{\theta}} - \sum_g^m \pi_g \theta_g = \sum_g^m \pi_g (Z_g' \Omega_g^{-1} Z_g)^{-1} (Z_g' \Omega_g^{-1} w_g) + o_p(1).$$

$$\hat{\hat{\theta}} - E[\theta_g] = \sum_g^m \pi_g (Z_g' \Omega_g^{-1} Z_g)^{-1} (Z_g' \Omega_g^{-1} w_g) + o_p(1).$$

Now, if we multiply by the stabilizing rate $\sqrt{N}$ and knowing that:

$$\sqrt{N}(\hat{\hat{\theta}} - E[\theta_g]) = \sqrt{N}\sum_g^m \pi_g (Z_g' \Omega_g^{-1} Z_g)^{-1} (Z_g' \Omega_g^{-1} w_g) + o_p(1).$$

In addition, we know that $N_g = N\pi_g$ such that $\sqrt{N} = \sqrt{\frac{N_g}{\pi_g}}$. Then,

$$\sqrt{N}(\hat{\hat{\theta}} - E[\theta_g]) = \sum_g^m \sqrt{\pi}_g \sqrt{N_g} (Z_g' \Omega_g^{-1} Z_g)^{-1} (Z_g' \Omega_g^{-1} w_g) + o_p(1).$$

By Lindeberg-Levy CLT we have that $\frac{1}{\sqrt{N_g}}(Z_g' \Omega_g^{-1} w_g)\xrightarrow{d}N(0, Q_g)$ with $Q_g = plimN_g^{-1}(Z_g' \Omega_g^{-1} Z_g)$. Then, $\sqrt{N_g}(Z_g' \Omega_g^{-1} Z_g)^{-1}(Z_g' \Omega_g^{-1} w_g)\xrightarrow{d}N(0, Q_g^{-1})$. Thus, we can conclude that:

$$\sqrt{N}(\hat{\hat{\theta}} - E[\theta_g])\xrightarrow{d}N(0, \sum_g^m \pi_g Q_g).$$

19.1.4. Proof Theorem 4

Proof. We known that the cluster-specific GLS estimators are equal to:

[7] I am thankful to Dr. Abhishek Ananth for the sketch of this proof.

$$\hat{\theta}_{g,GLS} = E[\theta_g] + \alpha_g + (Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g),$$

with:

$$w_g = diag(X_g)\lambda_g + \epsilon_g.$$

The presence of the lagged dependent variable in the left hand side of our model causes a bias of order $(n_g)^{-1}$ in $\hat{\theta}_{g,GLS}$:

$$E(\hat{\theta}_{GLS,g} - \theta_g) = K_{g,n_g} + O_p(n_g^{-3/2}) = \delta_{g,n_g}.$$

The derivation of the small-sample bias is presented in subsection 19.1.5.
Now, we can re-write the cluster specific estimator as:

$$\hat{\theta}_{g,GLS} = E[\theta_g] + \alpha_g + \delta_{g,n_g} + [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}],$$

with the difference in squared brackets of order n_g^{-1} because δ_{g,n_g} is $O_p(n_g^{-1})$.
Using the previous expression, we can write the Mean-Cluster estimator as :

$$\frac{1}{m}\sum_g \hat{\theta}_{g,GLS} = E[\theta_g] + \sum_g \frac{1}{m}\alpha_g + \frac{1}{m}\sum_g \delta_{g,n_g} + \frac{1}{m}\sum_g [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}].$$

From now on, we use a diagonal path asymptotic theory framework (Phillips and Moon (1999), Levin et al. (2002)) represented by $(m(n_g), n_g)_{diag} \to \infty$.
Since $\delta_{g,n_g} = O_p(n_g^{-1})$, $[(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}] = O_p(n_g^{-1})$ and the growth of m requires the growth of n_g because m is a monotonic increasing function of n_g, we have:

$$\underset{(m(n_g),n_g)_{diag}\to\infty}{plim} \frac{1}{m}\sum_g \delta_{g,n_g} = \underset{(m(n_g),n_g)_{diag}\to\infty}{plim} \frac{1}{m(n_g)}\sum_g \frac{n_g}{n_g} O_p(n_g^{-1}) = \underset{(m(n_g),n_g)_{diag}\to\infty}{plim} \frac{1}{m(n_g)}\sum_g \frac{1}{n_g} O_p(1) = 0,$$

Similarly,

$$\underset{(m(n_g),n_g)_{diag}\to\infty}{plim} \frac{1}{m(n_g)}\sum_g [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}] = 0.$$

Now, using $\sqrt{m}$ as the stabilizing factor, we obtain:

$$\sqrt{m}(\frac{1}{m}\sum_g \hat{\theta}_{g,GLS} - E[\theta_g]) = \frac{1}{\sqrt{m}}\sum_g \alpha_g + \frac{1}{\sqrt{m}}\sum_g \delta_{g,n_g} + \frac{1}{\sqrt{m}}\sum_g [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}].$$

We can re-write it as:

$$\begin{aligned}\sqrt{m}(\frac{1}{m}\sum_g \hat{\theta}_{g,GLS} - E[\theta_g]) &= \frac{1}{\sqrt{m}}\sum_g \alpha_g + \frac{\sqrt{m}}{m}\sum_g^m (K_{g,n_g} + O_p(n_g^{-3/2})) \\ &+ \frac{1}{\sqrt{m}}\sum_g [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}].\end{aligned} \tag{19.76}$$

Then,

$$\sqrt{m}(\frac{1}{m}\sum_g \hat{\theta}_{g,GLS} - E[\theta_g]) = \frac{1}{\sqrt{m}}\sum_g \alpha_g + \frac{\sqrt{m}}{m}\sum_g^m (O_p(n_g^{-1}) + O_p(n_g^{-3/2})) \\ + \frac{1}{\sqrt{m}}\sum_g [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}]. \tag{19.77}$$

$$\sqrt{m}(\frac{1}{m}\sum_g \hat{\theta}_{g,GLS} - E[\theta_g]) = \frac{1}{\sqrt{m}}\sum_g \alpha_g + \frac{1}{m}\sum_g^m \frac{\sqrt{m}}{n_g}(O_p(1) + O_p(n_g^{-1/2})) \\ + \frac{\sqrt{m}}{m}\sum_g [(Z_g'\Omega_g^{-1}Z_g)^{-1}(Z_g'\Omega_g^{-1}w_g) - \delta_{g,n_g}]. \tag{19.78}$$

The last two terms converge to 0 by Assumption 9.4.
Since m is a monotonic increasing function of n_g and by Linderberg-Levy CLT, we obtain:

$$\sqrt{m}(\frac{1}{m}\sum_g \hat{\theta}_{g,GLS} - E[\theta_g]) \xrightarrow{d} N(0, \Delta_\alpha).$$

19.1.5. Small Sample Bias of $\hat{\theta}_g$ In order to derive the bias of $\hat{\theta}_g$, I follow Kiviet and Phillips (1993) and Grubb and Symons (1987) and express the dependent variable for each individual as:

$$y_{gi} = \tilde{F}_g y_{gi0} + \tilde{C}_g X_{gi}\beta_g + \tilde{C}_g \tilde{X}_{gi}\lambda_{gi} + \varepsilon_{gi}, \tag{19.79}$$

where:

$$y_{gi} = \begin{bmatrix} y_{gi0} \\ y_{gi1} \\ y_{gi2} \\ y_{gi3} \\ \cdots \\ y_{giT-1} \end{bmatrix}, \tilde{F}_g = \begin{bmatrix} 1 \\ \rho_g \\ \rho_g^2 \\ \rho_g^3 \\ \cdots \\ \rho_g^{T-1} \end{bmatrix}, x_{gi} = \begin{bmatrix} x_{gi1} \\ x_{gi2} \\ x_{gi3} \\ x_{gi4} \\ \cdots \\ x_{giT} \end{bmatrix}, \tilde{C}_g = \begin{bmatrix} 0 & 0 & 0 & \cdots\cdot & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \rho_g & 1 & 0 & \cdots & 0 \\ \rho_g^2 & \rho_g & 1 & \cdots & 0 \\ \cdots\cdot\cdot & & & & \\ \rho_g^{T-1} & \rho_g^{T-2} & 1 & \cdots & 0 \end{bmatrix}, \lambda_{gi} = \begin{bmatrix} \lambda_{gi1} \\ \lambda_{gi2} \\ \lambda_{gi3} \\ \lambda_{gi4} \\ \cdots \\ \lambda_{giT} \end{bmatrix}, \varepsilon_{gi} = \begin{bmatrix} \varepsilon_{gi1} \\ \varepsilon_{gi2} \\ \varepsilon_{gi3} \\ \varepsilon_{gi4} \\ \cdots \\ \varepsilon_{giT} \end{bmatrix}.$$

If I stack up individual vectors in a group one, I obtain:

$$y_g = F_g y_{g0} + C_g x_g \beta_g + C_g \tilde{X}_g \lambda_g + \varepsilon_g, \tag{19.80}$$

with $F_g = diag(\tilde{F}_g)$, $C_g = diag(\tilde{C}_g)$, $\tilde{X}_g = diag(\tilde{X}_{gi})$.
Also, I know that the estimator per cluster is given by:

$$\hat{\theta} = (Z_g'\Omega^{-1}Z_g)^{-1}(Z_g'\Omega^{-1}y),$$

with $Z_g = [y_{g-1}X_g]$.
Now, I can define:

$$E[Z_g] = \bar{Z}_g + C_g u_g e_1',$$

where $\bar{Z}_g = [F_g y_{go} \quad C_g X_g]$.
Then, the bias of the estimator is given by:

$$E[\hat{\theta}_g - \theta_g] = E[(Z_g'\Omega^{-1}Z_g)^{-1}(Z_g'\Omega^{-1}u_g)],$$

with $Z_g = [y_{g-1} \quad X_g]$.

Following Kiviet and Phillips (1993), I find that:

$$\begin{aligned} E[\hat{\theta}_g - \theta_g] = -(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1}[\bar{Z}_g'\Omega_g^{-1}C_g\bar{Z}_g(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1}e_1 + \\ e_1 tr(C_g'\Omega_g^{-1}\bar{Z}_g(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1}\bar{Z}_g\Omega_g) + 2e_1 tr(C_g'\Omega_g^{-1}C_g\Omega_g)]. \end{aligned} \tag{19.81}$$

This can be rewritten as:

$$E[\hat{\theta}_g - \theta_g] = K_{g,n_g} + o_p(n_g^{-1}) = O_p(n_g^{-1}).$$

REFERENCES

Andrabi, T., J. Das, A. I. Khwaja, and T. Zajonc (2011). Do value-added estimates add value? accounting for learning dynamics. *American Economic Journal: Applied Economics 3*(3), 29–54.

Arellano, M. and S. Bond (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. *The review of economic studies 58*(2), 277–297.

Bai, J. (2013). Fixed-effects dynamic panel models, a factor analytical method. *Econometrica 81*(1), 285–314.

Bai, J. and K. Li (2021). Dynamic spatial panel data models with common shocks. *Journal of Econometrics 224*(1), 134–160. Annals Issue: PI Day.

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

Betancourt, M. J. and M. Girolami (2013). Hamiltonian monte carlo for hierarchical models.

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

Breitung, J., S. Kripfganz, and K. Hayakawa (2022). Bias-corrected method of moments estimators for dynamic panel data models. *Econometrics and Statistics 24*, 116–132.

Bun, M. J. and F. Windmeijer (2010). The weak instrument problem of the system gmm estimator in dynamic panel data models. *The Econometrics Journal 13*(1), 95–126.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics 90*(3), 414–427.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of human resources 50*(2), 317–372.

Chamberlain, G. (1979). Analysis of covariance with qualitative data.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics 34*(3), 305–334.

Choi, I. (2015). Panel cointegration. *The Oxford handbook of panel data*.

Chudik, A. and M. H. Pesaran (2022). An augmented anderson–hsiao estimator for dynamic short-t panels. *Econometric Reviews 41*(4), 416–447.

Cule, E. and M. De Iorio (2012). A semi-automatic method to guide the choice of ridge parameter in ridge regression. *arXiv preprint arXiv:1205.0686*.

Dynan, K. E. (2000). Habit formation in consumer preferences: Evidence from panel data. *American Economic Review 90*(3), 391–406.

Frühwirth-Schnatter, S. and R. Tüchler (2008). Bayesian parsimonious covariance estimation for hierarchical linear mixed models. *Statistics and Computing 18*(1), 1–13.

Greene, W. H. (2008). The econometric approach to efficiency analysis. *The measurement of productive efficiency and productivity growth 1*(1), 92–250.

Grubb, D. and J. Symons (1987). Bias in regressions with a lagged dependent variable. *Econometric Theory 3*(3), 371–386.

Hahn, J. and W. Newey (2004). Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica 72*(4), 1295–1319.

Hamilton, J. D. (1994). *Time series analysis*. Princeton university press.

Hansen, B. E. and S. Lee (2019). Asymptotic theory for clustered samples. *Journal of econometrics 210*(2), 268–290.

Hausman, J. A. and W. E. Taylor (1981). Panel data and unobservable individual effects. *Econometrica: Journal of the Econometric society*, 1377–1398.

Heckman, J. J. (1987). *The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process and some Monte Carlo evidence*. University of Chicago Center for Mathematical studies in Business and Economics.

Hoerl, A. E., R. W. Kannard, and K. F. Baldwin (1975). Ridge regression: some simulations. *Communications in Statistics-Theory and Methods 4*(2), 105–123.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge university press.

Hsiao, C. (2020). Estimation of fixed effects dynamic panel data models: linear differencing or conditional expectation. *Econometric Reviews 39*(8), 858–874.

Hsiao, C., M. Hashem Pesaran, and A. Kamil Tahmiscioglu (2002). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. *Journal of Econometrics 109*(1), 107–150.

Hsiao, C., Q. Li, Z. Liang, and W. Xie (2019). Panel data estimation for correlated random coefficients models. *Econometrics 7*(1).

Hsiao, C., D. C. Mountain, M. L. Chan, and K. Y. Tsui (1989). Modeling ontario regional electricity system demand using a mixed fixed and random coefficients approach. *Regional Science and Urban Economics 19*(4), 565–587.

Hsiao, C., M. H. Pesaran, and A. K. Tahmiscioglu (1998). Bayes Estimation of Short-run Coefficients in Dynamic Panel Data Models.

Im, K. S., M. H. Pesaran, and Y. Shin (2003). Testing for unit roots in heterogeneous panels. *Journal of econometrics 115*(1), 53–74.

Kapetanios, G., C. Mastromarco, L. Serlenga, and Y. Shin (2017). Modelling in the presence of cross-sectional error dependence. In *The Econometrics of Multi-dimensional Panels*, pp. 291–322. Springer.

Kiviet, J. F. and G. D. A. Phillips (1993). Alternative bias approximations in regressions with a lagged-dependent variable. *Econometric Theory 9*(1), 62–80.

Krishnakumar, J., M. Avila Márquez, and L. Balazsi (2017). *Random Coefficients Models*, pp. 125–161. Springer International Publishing.

Levin, A., C.-F. Lin, and C.-S. J. Chu (2002). Unit root tests in panel data: asymptotic and finite-sample properties. *Journal of econometrics 108*(1), 1–24.

Matyas, L. (2017). *The econometrics of multi-dimensional panels*. Springer.

Moon, H. R., M. Shum, and M. Weidner (2018). Estimation of random coefficients logit demand models with interactive fixed effects. *Journal of Econometrics 206*(2), 613–644. Special issue on Advances in Econometric Theory: Essays in honor of Takeshi Amemiya.

Moreira, M. J. (2009). A maximum likelihood method for the incidental parameter problem. *The Annals of Statistics 37*(6A), 3660 – 3696.

Mundlak, Y. (1961). Aggregation over time in distributed lag models. *International Economic Review 2*(2), 154–163.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, 69–85.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica 49*(6), 1417–1426.

Pesaran, M. and R. Smith (1995). Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics 68*(1), 79 – 113.

Pesaran, M. H., Y. Shin, and R. P. Smith (1999). Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association 94*(446), 621–634.

Phillips, P. C. and H. R. Moon (1999). Linear regression limit theory for nonstationary panel data. *Econometrica 67*(5), 1057–1111.

Phillips, R. F. (2010). Iterated feasible generalized least-squares estimation of augmented dynamic panel data models. *Journal of Business & Economic Statistics 28*(3), 410–422.

Rendon, S. R. (2013). Fixed and random effects in classical and bayesian regression. *Oxford Bulletin of Economics and Statistics 75*(3), 460–476.

Rossi, P. and G. Allenby (2009). Bayesian applications in marketing. In *The Oxford Handbook of Bayesian Econometrics*. Citeseer.

Sarafidis, V. and D. Robertson (2009). On the impact of error cross-sectional dependence in short dynamic panel estimation. *The Econometrics Journal 12*(1), 62–81.

Sarafidis, V. and T. Wansbeek (2021). Celebrating 40 years of panel data analysis: Past, present and future.

Sims, C. A. (2000). Using a likelihood perspective to sharpen econometric discourse: Three examples. *Journal of Econometrics 95*(2), 443 – 462.

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

Wooldridge, J. M. (2003). Cluster-sample methods in applied econometrics. *American Economic Review 93*(2), 133–138.

Wooldridge, J. M. (2005a). Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Review of Economics and Statistics 87*(2), 385–390.

Wooldridge, J. M. (2005b). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics 20*(1), 39–54.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

ACKNOWLEDGEMENTS

I am grateful to Prof. Hashem Peasaran, Prof. Jeffrey Wooldridge, Prof. Aleksey Tetenov, Prof. Jaya Krishnakumar, Prof. Stefan Sperlich, Prof. Domenico Giannone, Prof. Sebastian Engelke, Prof. Sylvia Fruhwirth-Schnatter, Prof. Giacomo De Giorgi, Asst.Prof. Abhishek Ananth, Prof. Jörg Breitung.