pp. 1–45.

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

Monika Avila Márquez[†]

[†] *Institute of Economics and Econometrics, Geneva School of Economics and Management, University of Geneva, Uni Mail, Boulevard du Pont d'Arve 40, Geneva 1205, Switzerland.*
E-mail: monika.avila@unige.ch

Summary This paper investigates the identification and estimation of dynamic heterogeneous linear models for unbalanced panel data with known clustering structure and short time dimension (greater than or equal to 3). For this purpose, I use a linear multidimensional 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. For estimation of the mean coefficients, I propose a Mean Cluster-FGLS estimator and a Mean Cluster-OLS estimator. In order to make feasible the GLS estimation of the cluster specific parameters, I introduce a ridge estimator of the variance-covariance matrix of the model. The Mean Cluster estimators are consistent when: i) 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 when ii) the number of clusters grows at a slower rate than the growth rate of the number individuals per cluster. In addition, I present two extensions of the baseline model. In the first one, I allow for cluster-individual specific fixed effects instead of cluster additive fixed effects. In this setting, I propose a Hierarchical Bayes estimator that takes into account the problem of unknown initial conditions. In the second extension, I allow for cross sectional dependence by including common factors. For estimation of this model, I propose the Mean Cluster estimator using the time demeaned variables. As an empirical application, I present the estimation of a value-added model of learning.

Keywords: *Panel Data, Clustering, Mixed Coefficients, Variance-Covariance estimation, Ridge, Bayesian Analysis.*

1. INTRODUCTION

Heterogeneous linear dynamic panel data models with short time dimension (T) suffer of two well-known problems: the incidental parameter bias (Nickell, 1981) and the unknown initial conditions dependency (Hsiao, 2020, Wooldridge, 2005b). When the slope coefficients are heterogeneous across individuals and the time dimension is equal to 3, GMM estimation is unfeasible. Similarly debiased Mean Group (MG) estimation (Pesaran and Smith, 1995) is unfeasible since available debiasing techniques, jackknife (Dhaene and Jochmans, 2015) and analytical (Kiviet and Phillips, 1993), are suitable only for T larger than 3. For short time dimensions greater than 3, debiasing is possible but it could be negatively affected by poor first stage estimates. [1] Another limitation of mean group estimation is that the small number of time observations prevents the inclusion of a big number of covariates.

While estimation of dynamic linear panel data models with unobserved multiplicative

[1] Iterative estimation is helpful for homogeneous panel data (Hahn and Newey, 2004).

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 method for dynamic linear heterogeneous panel data models that exploits the clustering structure in the data. For this purpose, I assume that individual unobserved heterogeneity is partitioned into two components: individual heterogeneity correlated with the regressors that is pooled within clusters and individual heterogeneity that is uncorrelated with the regressors within clusters. Under this key assumption, it is possible to obtain consistent estimates and overcome the incidental parameter bias as well as the initial conditions problem.

More specifically, 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 while the former assumption states that the coefficient vector is equal to 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 the inclusion of 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. Another advantage is that the initial condition dependency is controlled. In contrast, a disadvantage of the inclusion of cluster fixed effects instead of individual effects is that the model is misspecified if individuals do not pool within clusters. I address this problem by extending the model and allowing for additive cluster-individual fixed effects. Another problem surges if the assumed clustering structure is not correct.

In particular, I 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, I use a three dimensional panel data framework and consider the following baseline model for individual i belonging to cluster g:

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

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. [2] The number of groups in the panel is equal to m, the number of individuals per group is equal to N_g and the number of observations per

[2] I could have used the alternative notation i_g that represents individual i belonging to cluster g and t_{i_g} for time observation t of individual i belonging to group g as explained in Section 2. However, I use three indexes in order to simplify the notation.

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. [3] Additionally, I 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 mean.

For estimation of the baseline model 1.1, I propose two Mean Cluster (MC) estimators and the cluster specific estimators. The Mean Cluster estimators are consistent when: i) 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 when ii) the number of clusters grows at a slower rate than the growth rate of the number individuals per cluster.

The Mean Cluster estimators are the mean of the FGLS or OLS parameter estimations of each cluster g. In order to make feasible GLS, I propose a ridge estimation of the variance-covariance components along with a modification suitable for big sample size. 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 estimators are more efficient because they use a larger sample size, and v) 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 disadvantage of the Mean Cluster FGLS or OLS estimators are i) not robust to violation of cluster assumption and ii) instability when the proportion of observed clusters is too small.

In order to test the assumption of clustered individual heterogeneity, I propose two specification tests that are simple extensions of Hausman test (Hausman and Taylor, 1981). I propose to compare the Mean Cluster estimator with the simple Pooled OLS estimator in order to test the null hypothesis of homogeneity versus heterogeneity. In case of rejection of the hypothesis of homogeneity, one can test for the presence of individual fixed effects by comparing the Mean Cluster estimator against any other GMM, MG 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, I extend the baseline model 1.1 to allow for the presence of cluster-individual specific additive effects. For this setting, I propose a bayesian hierarchical estimator. Another issue is the ignorance of cross sectional dependence. In order to deal with this problem, I extend the baseline model 1.1 to a setting that includes common factors and I propose Mean Cluster estimation using the time demeaned variables (Sarafidis and Robertson (2009)).

[3] Model 1 can be rewritten using a two dimensional panel data model for individual i:

$$y_{it} = \alpha_i + \rho_i y_{it-1} + x'_{it}\beta_{it} + \varepsilon_{it} \qquad \forall t \in (1, 2, ..., T_i).$$

In the two-level setting for micro-panel data, parameters can be treated as random variables since individuals can be considered as random draws from a common population (Wooldridge, 2010). Consequently, β_{it} can be considered as random vector defined on a probability space with sigma-algebra $\mathcal{F}$. Then, the cluster specific coefficients are equal to the conditional expectation of β_{it} on the cluster specific-sigma field ($\mathcal{F}_g$) contained in $\mathcal{F}$ ($\beta_g = E[\beta_{it}|\mathcal{F}_g]$). Similarly, $\rho_g = E[\rho_i|\mathcal{G}_g]$ and $\alpha_g = E[\alpha_i|\mathcal{H}_g]$ where $\mathcal{G}_g$ and $\mathcal{H}_g$ are cluster specific sigma fields contained into $\mathcal{G}$ and $\mathcal{H}$.

As an empirical application, I present the estimation of a value-added model of learning. This illustrates the use of the Mean Cluster estimator in a model that includes cluster individual specific effects, cluster common factors, and autocorrelated disturbance terms. Andrabi et al. (2011) show the importance of persistence in program evaluation. They estimate the impact of enrollment in private school using a value-added model of learning with panel data obtained by means of cluster sampling of villages in three districts of Pakistan. The authors faced three main empirical challenges: i) the time dimension is equal to 3, ii) the dependent variables; scores of Mathematics, Urdu, and English; present measurement error, and iii) the presence of individual specific unobserved heterogeneity. In order to eliminate the individual specific effects, the authors use the first-differenced model. But due to the short time dimension and the measurement error, this transformation leaves them with no available lags of the dependent variable that can be used as instruments for GMM estimation. As a solution, the authors assume that the measurement error is uncorrelated across subjects and they use as instrumental variables the score of other subjects. But a failure of this key assumption invalidates their identification strategy. In order to propose an alternative estimation procedure, I use the Mean Cluster estimator after time demeaning the variables and I use the lag of the time demeaned regressors as instrumental variables. The estimated average persistence parameter is 0.64, and the average effect of private school on scores is 0.30. The estimated persistence parameter is larger than the one presented by Andrabi et al. (2011) (0.10) and the estimated effect of private school is smaller than the one presented by the authors (0.42).

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 in panels with long time dimension and suggests corrections when time dimension is short. Bester and Hansen (2016) propose a grouped estimator for fixed effects non-linear models based on observable characteristics. Bonhomme and Manresa (2015) propose a grouping algorithm to classify individuals based on observables and unobservables.

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 [4], ii) it proposes a Mean Cluster estimator, appropriate for settings when debiasing is not feasible, iii) it provides the conditions for consistency of the Mean Cluster estimators, iv) it provides an estimation method for the variance-covariance of the model by extending the method presented by Krishnakumar et al. (2017) to a dynamic setting, and v) it proposes a hierarchical Bayesian estimator that takes into account the initial conditions.

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

[4] This assumption is not equal to the one proposed by Hsiao et al. (1989). The authors proposed a mixed fixed and random coefficients framework which means that some regressors present fixed coefficients and other random coefficients while 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 and the necessary assumptions, Section 4 states the identification of the parameters of interest, Section 5 presents the estimation strategy, Section 6 exposes the statistical properties of the methods proposed, Section 7 explains possible limitations of model 1.1, Section 8 presents specification tests, Section 9 relaxes the assumption of additive cluster effects to cluster-individual additive specific effects and presents a Hierarchical Bayes estimator, Section 10 presents an extension of the model with cross sectional dependence, Section 11 describes the Monte Carlo experiments and the results, Section 12 presents an empirical application, Section 13 gives the conclusions.

Notation: $||\cdot||^2$ is the Euclidean norm. $||\cdot||_F$ is the Frobenius norm. Scalar random variables are collected in row vectors, for instance y_{git} can be collected in the the vector $Y \in \mathbb{R}^M$ ($Y = (y_{111}, ..., y_{mN_mT_{i_m}})$). The transpose of random column vectors are collected in matrices, for instance K regressors x_{git} are stack up in the matrix $X \in \mathbb{R}^{M\times K}$. I_A represents the identity matrix with dimension $A \times A$ where A is a positive integer.

2. DATA STRUCTURE

The data $\{y_{it}, x_{it}\}_{i=1}^{N}$ is obtained from stratified sampling and it can be partitioned in non overlapping subsets $\{y_{git}, x_{git}\}_{gi=1}^{N_g}$. This means that the population is stratified in m non overlapping independent known clusters. For each cluster g, N_g individuals are sampled over T_{i_g} periods of time. 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 $M = \sum_g^m n_g$. This data can be seen as unbalanced three level panel.

I 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, I use i and t equivalently to i_g and t_{i_g}. This does not mean that I assume that individual i is not subordinated to g.

3. THE MODEL

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

$$y_{git} = \alpha_{1,g} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git} \qquad t = 1, ..., T_{i_g}, \tag{3.1}$$

with:

$$\beta_{git} = \beta_g + \lambda_{git}. \tag{3.2}$$

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 $E[\rho_g]$, $E[\beta_g]$. [5]

The total number of time observations per individual T_{i_g} is small and considered as 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. This setting can be evaluated using an asymptotic sequence framework where I allow N_g to grow but 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 just allow for additive individual fixed effects. In this paper, I handle this problem by imposing clustered heterogeneity and using a novel mixed structure in the slope coefficients.

More specifically, I 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, I 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 the heterogeneous habit formation of individuals in a certain cluster is driven by common cultural unobserved characteristics while possible deviations are random and non correlated to "taste-shifters". [6]

This coefficient structure can have two possible interpretations: i) the data is sampled from a density function with heterogeneous parameters or ii) the regressors are freely correlated to cluster specific effects while preserving non correlation with cluster-individual-time specific effects. The latter could be the possible if one is willing to assume that the correlation of the regressors with unobserved individual heterogeneity is equal within clusters. For instance, the inner ability and the marginal return to education of individuals is equally correlated to education within a city if we believe that individuals with higher ability do not only self-select into education levels but also into the city where they will have the highest return to their education (See Appendix C).

This model is relevant for different empirical applications since it permits to account for correlated cluster heterogeneity as well as individual and time heterogeneity. For instance, one could be interested in the the study of dynamic heterogeneous demand equations, the heterogeneity of habit formation, the heterogeneity of income persistence, the dynamic heterogeneous treatment effects and others.

In the following lines, I present the assumptions of the model with more detail.

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

The clusters are known by the researcher based on observed characteristics. For instance, clustering can be done by counties or sub-regions, economic activity categories at a

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

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

detailed level, etc. 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 important to notice that cluster belonging does not vary with time.

REMARK 3.1. 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 be the case for households within sub-region or enterprises in an economic sector.

ASSUMPTION 3.3. *Non vanishing clusters.*

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

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.

REMARK 3.2. This assumption implies that the number of clusters is fixed.

REMARK 3.3. It is possible to assume that the number of clusters grows. In this case, it is necessary to add a restriction to its growth rate by assuming that it grows at a slower rate than the number of individuals in the cluster such as $\frac{\sqrt{m(n_g)}}{n_g} \to 0$. 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 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 method given in section 5 is still consistent for both cluster and mean coefficients. Nevertheless, the asymptotic framework is different from the one presented in section 6 and it is provided in the Appendix C.

REMARK 3.4. 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) is equal to 1.*

This assumption is line with a setting where the available sample is obtained from

stratified sampling. Abadie et al. (2017) discusses the importance of the proportion of observed clusters.

Remark 3.5. The proportion of observed clusters can be lower than 1. This means that not all clusters are sampled and as a result one can assume that the data available is obtained from cluster sampling. In the Appendix C, I present the assumptions that are compatible with this setting. An example of this data is the one used by Andrabi et al. (2011) and that I employ for the empirical application in Section 12.

Assumption 3.5. *Fixed cluster specific persistence parameter.*

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

$$\alpha_2, g = \rho_g - E[\rho_g].$$

Assumption 3.6. *Fixed cluster additive specific effects* $\alpha_{1,g}$.

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

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

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

$$\alpha_3, g = \beta_g - E[\beta_g].$$

The unobserved coefficient vector is composed of a fixed cluster 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 3.8. y_{git} *are generated from the stationary process with initialization values* $y_{gi,-h_{i_g}}$ *sampled* h_{i_g} *number of periods before the data collection in period 0.*

This implies that the initial observations are given by:

$$y_{gi0} = \rho_g^{h_{i_g}} y_{gi,-h_{i_g}} + \alpha_{1,g}\frac{1-\rho_g^{h_{i_g}}}{1-\rho_g} + \sum_{l=0}^{h_{g_i}} \rho_g^l x'_{gi-l}\beta_{gi-l} + \sum_{l=0}^{h_{i_g}} \rho_g^l \varepsilon_{gi-l}. \tag{3.3}$$

h_{i_g} is set free, this is possible thanks to the Assumption 3.6. If the model presents cluster-individual additive fixed effects instead of cluster additive effects and h_{i_g} is small, the individual initialization values are important. In that case, there is need to add an additional 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. This is similar to Hsiao et al. (2002).

Assumption 3.9. 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 to the one presented by Hsiao et al. (2002).

REMARK 3.6. The method presented in section 5 allows for trend stationary regressors only if the data generating process started a short time ago (small h_{i_g}). An example could be the wage of young individuals and in this case, one can include age and experience as regressors in our model.

REMARK 3.7. This assumption states that the dependent variable and the regressors are both integrated of order 0. Additionally, it is necessary when the model presents cluster-individual additive specific effects (Assumption 9.1) instead of cluster additive specific fixed effects (Assumption 3.6).

REMARK 3.8. Under this assumption, 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.

ASSUMPTION 3.10. *The random cluster-individual-time effects are zero mean conditional on the covariates.*

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

This implies that $E[\beta_{git}|x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}] = \beta_g$.

ASSUMPTION 3.11. *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 that the model presents dynamic completeness without conditioning on cluster effects because cluster specific effects are considered as fixed parameters. However, one can also condition on cluster specific effects and obtain the same orthogonality conditions presented in section 4 if one would like to assume for correlated cluster random effects (See Appendix C).

REMARK 3.9. 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. This assumption 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 in 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 5 by replacing OLS and GLS by GMM using instrumental variables. In the empirical application, I use GMM estimation with instrumental variables.

ASSUMPTION 3.12. *Error term ε_{git} is identically and independently distributed over t and i in each cluster g but heteroskedastic across clusters.*

$$E[\varepsilon_{git}] = 0, \quad E[\varepsilon_{git}^2] = \sigma_{\varepsilon_g}^2 < \infty.$$

4. IDENTIFICATION

We can rewrite the model as:

$$y_{git} = \rho_g y_{git-1} + \alpha_{1,g} + x'_{git}\beta_g + u_{git} = z'_{git}\theta_g + u_{git}, \tag{4.1}$$

where: $z_{git} = [y_{git-1}, 1, x'_{git}]'$, $\theta_g = [\rho_g, \alpha_{1,g}, \beta'_g]'$, $u_{git} = x'_{git}\lambda_{git} + \varepsilon_{git}$ is a composite error term.
Assumptions 3.10 and 3.11 imply the following orthogonality conditions: [7]

$$E[u_{git}x_{gis}] = 0 \quad \forall s \in (1,2,...,T), i \in (1,2,...,N_g), g \in (1,2,...,m), \tag{4.2}$$

$$E[u_{git}y_{git-1}] = 0 \quad \forall t \in (1,2,...,T), i \in (1,2,...,N_g), g \in (1,2,...,m). \tag{4.3}$$

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

$$E[u_{git}z_{git}] = 0 \quad \forall t \in (1,2,...,T), i \in (1,2,...,N_g). \tag{4.4}$$

Note that I 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, I also assume that the z_{git} is full rank.

ASSUMPTION 4.1. *The matrix $E[z_{git}z'_{git}]$ is full rank.*

5. ESTIMATION

If I rewrite the model 1.1 using backward substitution, I obtain the following expression of the dependent regressor:

$$y_{git} = \rho_g^t y_{gi0} + \sum_{l=0}^{t} \rho_g^l(\alpha_{1,g} + x'_{git-l}(\beta_g + \lambda_{git-l})) + \sum_{l=0}^{t}(\rho_g^l)\varepsilon_{git-l}. \tag{5.1}$$

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

$$y_{git-1} = \rho_g^{t-1} y_{gi0} + \sum_{l=0}^{t-1} \rho_g^l(\alpha_{1,g} + x'_{git-1-l}(\beta_g + \lambda_{git-1-l})) + \sum_{l=0}^{t-1}(\rho_g^l)\varepsilon_{git-1-l}. \tag{5.2}$$

It is easy to see from (5.2) that a GMM estimation ignoring the clustering structure of the data leads to inconsistent estimates of the mean parameters. This is caused by

[7] 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.

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. [8]

Similarly, one could argue that the researcher could perform Mean Group estimation per individual within cluster. This approach permits the estimation of the mean coefficient and could be used for estimation of 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. Thus, when the time dimension is fixed and the number of clusters is big it would be beneficial to use another estimation strategy.

In order to fill this gap, I propose a method that allows to estimate mean cluster and 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 cluster g:

$$E[u_{git} z_{git}] = 0 \quad \forall t \in (1, 2, ..., T_{i_g}). \tag{5.3}$$

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

$$\frac{1}{N_g} u_g' Z_g = 0 \quad \forall g \in (1, 2, ..., m). \tag{5.4}$$

It is easy to see that using the sample moment conditions 5.4 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-homoskedastic and non independent error term. 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 $\Omega_g = E[u_g u_g'] = diag(X_g)(I_K \otimes \Delta_{\lambda g}) diag(X_g) + \sigma^2_{\varepsilon_g} I_{N_g}$.
I propose an estimation procedure for the unknown Ω_g in Subsection 5.1.

Second stage: In the second stage, it is necessary to take the weighted mean of all estimated parameters ending up with a Mean Cluster estimator given by:

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

where $\hat{w}_g$ is an appropriate estimator of the importance of the cluster in the population, $\hat{\bar{\theta}}_{MC} = [\hat{\bar{\rho}}, \hat{\bar{\beta}}]$, $\hat{\theta}_g = [\hat{\rho}_g, \hat{\beta}_g]$.
I propose a weighted average of the cluster specific coefficients where the weights represent the importance of each cluster in the population.

[8] Ignoring cluster effects is equivalent to performing GMM estimation on the model: $\Delta y_{it} = \rho \Delta y_{it-1} + \Delta x_{it}' \beta + \Delta u_{it}$ with: $\Delta u_{it} = \Delta y_{it-1} \alpha_{2,g} + \Delta x_{it}' \alpha_{3,g} + \Delta x_{it}' \Delta \lambda_{git} + \Delta \varepsilon_{it}$, $\alpha_{2,g} = \rho_g - E[\rho_g]$ and $\alpha_{3,g} = \beta_g - E[\beta_g]$. Thus, we would not have available instruments.

REMARK 5.1. 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 while the MC averages cluster pooled estimators.

REMARK 5.2. In case of endogenous regressors, it is possible to replace the first-stage estimation with GMM estimation using instrumental variables. 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 needed to assume that the number of instrumental variables is equal or larger than the endogenous regressors. Finally, it is well known that OLS and FGLS estimation are special cases of GMM estimation using the regressors as their own instruments $p_{git} = x_{git}$

REMARK 5.3. The assumption of unobserved additive and multiplicative cluster fixed effects allows to estimate the specific parameters by pooling observations within each cluster. Additionally, OLS or GLS estimation is consistent under the assumptions presented in section 3 because the model is dynamic complete conditional con cluster specific effects.

REMARK 5.4. The Mean Cluster estimator is also consistent in a setting where the proportion of observed clusters is lower than 1 and the number of clusters grows at a slower rate than the number of individuals in the cluster. In this setting, one can assign an equal weight to all observed clusters. The assumptions for this setting is presented in Appendix C as well as the derivation of the statistical properties.

REMARK 5.5. Another possibility could be GMM estimation on the model in first differences using multiplicative cluster dummies when $T > 2$. But one could run into issues related to weak IVs (Bun and Windmeijer, 2010).

5.1. Variance-Covariance Estimation

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

$$\Omega_g = \sum_{k=1}^{K}\sum_{k'=1}^{K} \sigma_{\lambda_g,kk'} H_{g,kk',\lambda_g} + \sigma^2_{\epsilon_g} I_{n_g}. \tag{5.5}$$

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, it is necessary to obtain a first stage estimator of the residuals for each cluster which can be obtained using OLS estimation $r_{g_{OLS}} = (I_{n_g} - Z_g(Z'_gZ_g)^{-1}Z'_g)y_g = M_g w_g$ where $Z_g \in \mathbb{R}^{n_g x(K+1)}$ is the matrix stacking up all the observations for $z_{git} = [y_{git-1} \quad x'_{git}]'$.
Then, it follows that:

$$E[r_{g_{OLS}} r'_{g_{OLS}}] = M_g \Omega_g M_g. \tag{5.6}$$

Replacing expression (5.5) into equation (5.6) and applying the vec operator I obtain:

$$vec(E[r_{g_{OLS}} r'_{g_{OLS}}]) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \sigma_{\lambda_g, kk'} vec(M_g H_{g,kk',\lambda_g} M_g) + \sigma^2_{\epsilon_g} vec(M_g). \tag{5.7}$$

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

$$vec(E[r_{g_{OLS}} r'_{g_{OLS}}]) = B_{\lambda_g} vec(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(M_g). \tag{5.8}$$

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

$$vec(E[r_{g_{OLS}} r'_{g_{OLS}}]) = B_{\lambda_g} Dvech(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(M_g). \tag{5.9}$$

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

$$vec((r_{g_{OLS}} r'_{g_{OLS}})) = B_{\lambda_g} Dvech(\triangle_{\lambda_g}) + \sigma^2_{\epsilon_g} vec(M_g) + \nu_g. \tag{5.10}$$

Finally, notice that 5.10 is a simple linear model that can be rewritten as:

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

where:

$$R_g = vec(r_{g_{OLS}} r'_{g_{OLS}}),$$

$$C_g = [\quad B_{\lambda_g} D \quad vec(M_g)],$$

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

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

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

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

with $\tau \in [0, 2min(\zeta_{gl})]$ where ζ_{gl} is the eigenvalue l of the matrix $C'_g C_g$.
The penalization term using the l_2-norm allows to tackle the problem of high multicollinearity in the matrix $C'_g C_g$.

C. Large and Huge Sample Size

When the sample size is big, there are problems due to memory requirements for storage of vectorized matrices. In order to tackle this issue and reduce the computing requirements by half, I 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.

$$R_g = vech(r_g r_g'),$$

$$C_g = [B_{\lambda,g}, vech(M_g)],$$

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

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

6. STATISTICAL PROPERTIES

In this section, I present the statistical properties of the cluster specific estimators, the Mean Cluster estimator and the variance-covariance estimators using sequential asymptotic 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, I 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 3.4 to 3.11 and 4.1 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 3.4 to 3.11 and 4.1 hold, ii)* $\lim_{n_g \to \infty} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} C_{git} C_{git}' = M_g$ *with* $||M_g||_F < \infty$*. iii)* $\nu_{git} \sim iid(0, \sigma_\nu^2)$*, iv)* $\lim_{n_g \to \infty} \sum_{i_g}^{N_g} \sum_{t_{i_g}}^{T_{i_g}} C_{git} R_{git} = 0$*, 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*

$$\hat{\bar{\theta}} - \bar{\theta} \sim N(0, Q),$$

where $Q = \sum_g w_g^2 Q_g$.

7. ARE CLUSTER EFFECTS ENOUGH?

The estimator proposed in subsection 5 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 would be 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 cluster fixed 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, I present specification tests in the following section. I also present an extension of model 1.1 that includes cluster-individual additive effects.

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

8. SPECIFICATION TESTS

In this section, I present different specification tests to check for the presence of cluster specific effects.

First, we can check if cluster effects are enough to capture the heterogeneity in the panel. If the number of time observations is equal to 3 or greater, I propose a Hausman type (Hausman and Taylor (1981), Hsiao and Pesaran (2008)) test that compares a GMM estimator of the mean parameters with the Mean Cluster estimator proposed.

More specifically, the null and alternative hypothesis are the following:

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

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

The statistic is given by:

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

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

If the time dimension is larger than 3, one could replace the GMM estimator with the MG estimator.

Similarly, testing for cluster heterogeneity could be done with a Hausman type test that compares Pooled OLS estimator vs. a Mean Cluster estimator.

A study of the properties of the proposed tests is left for further research.

9. RELAXING THE ASSUMPTION OF CLUSTER ADDITIVE SPECIFIC EFFECTS

In this subsection, I relax the assumption of additive cluster specific effects and allow for the presence of additive cluster-individual correlated random effects. Therefore, Assumption 3.6 is replaced by the following one:

ASSUMPTION 9.1. *Correlated cluster-individual additive specific random effects* $\alpha_{1,gi}$.

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

$$y_{git} = \alpha_{1,gi} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}, \tag{9.1}$$

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

The estimation of model 9.1 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, I use a mean conditional approach instead of a linear difference approach. I 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 $\mathbb{E}(\alpha_{gi}|x_{it})$ 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, I re-express $\alpha_{1,gi}$ as a linear projection on the individual means of the regressors:

$$\alpha_{1,gi} = \bar{x}'_{gi.}\pi_g + \upsilon_{gi}, \tag{9.2}$$

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

This linear projection can be replaced in model 9.1:

$$y_{git} = \alpha_{1,gi} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}. \tag{9.3}$$

Now, it is only left the issue of the unobserved initial conditions. If I assume that the initial conditions are generated from the long-term mean, I can write them as:

$$y_{gi0} = \frac{\alpha_{gi}}{1-\rho_g} + \varepsilon_0. \tag{9.4}$$

Now, I can replace the linear projection of the individual effects on the individual mean of the regressors to obtain:

$$y_{gi0} = \frac{\bar{x}'_{gi.}\pi_g}{1-\rho_g} + \frac{\upsilon_{gi}}{1-\rho_g} + \varepsilon_0. \tag{9.5}$$

The combination of 9.3 and 9.5 leads to the system of equations:

$$\begin{aligned} y_{git} &= \bar{x}'_{gi.}\pi_g + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon^*_{git}, \quad t = 1, ..., T_{i_g}, \\ y_{gi0} &= \frac{\bar{x}'_{gi.}\pi_g}{1-\rho_g} + \frac{\upsilon_{gi}}{1-\rho_g} + \varepsilon_0. \end{aligned} \tag{9.6}$$

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

The likelihood of the observed data is given by:

$$L_{\zeta|y,y_{-1},X} = \prod_{g}^{m} \prod_{i}^{N_g} L(\zeta_g|y_{gi}), \tag{9.7}$$

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

The prior distributions for the fixed parameters are:

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

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

$$(\pi_g|\pi) \sim N(\pi, FF').$$

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 location parameter equal to 0.5 and 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 low correlation.

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

$$\beta_{git} = \beta_g + H_{\Delta_\lambda} z_{git}, \tag{9.8}$$

where z_{git} is a multivariate standard normal variable and $H_{\Delta_{\alpha,2}}$ 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.

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, I relax it and set up the following prior distribution for the initial conditions:

$$f_{y_{gi0}} \sim N(\mu_{0,gi}, \sigma^2_{y_0}), \tag{9.9}$$

where $\mu_{0,gi} = \frac{\alpha_{gi}}{1-\rho_g}$ 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 9.1. 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.

REMARK 9.2. 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.

9.1. INITIAL CONDITIONS: A MORE FLEXIBLE PRIOR?

A failure in the assumption of the DGP of y_{gi0} causes invalid inference and inconsistent estimates. In order to relax the assumptions 3.8 and 3.9, I 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 3.9, 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 form the stationary process.

In order to propose a prior that does not constraint the unconditional stationary distribution, I propose the following joint prior:

$$\begin{pmatrix} y_0 \\ \\ \theta \end{pmatrix} \sim N \left(\begin{pmatrix} \iota_{mNT} \otimes \mu_y \\ \\ \iota_{mNT} \otimes \bar{\theta} \end{pmatrix} , \ \Sigma_{y,\theta} \right),$$

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

10. HOW ABOUT CROSS-SECTIONAL DEPENDENCE?

10.1. A model including common global factors

The models 1.1 and 9.1 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 modelled either using spatial or factor models or a combination of both.

In this section, I extend model 1.1 in order to allow for cross-sectional dependence using a factor model. For this purpose, I 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 $\sum_{i_g}^{N_g} \Lambda'_{i_g} f_{gt}$ with factor loadings equal to 0 or 1 (Bonhomme and Manresa (2015), Kapetanios et al. (2017), Bai and Li (2021)).

Additionally, I include time specific effects that capture correlation across clusters. Once more following Bonhomme and Manresa (2015) and Kapetanios et al. (2017), it is possible to re-express time specific effects as $\gamma_t = \sum_i^N F'_g f_t$ with F_g representing factor loading g that takes value 0 or 1.

The extended model 1.1 includes cluster-time additive effects as well as time fixed effects as common factors:

$$y_{git} = \alpha_g + \gamma_t + \tau_{gt} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \varepsilon_{git}, \quad t = 1, ..., T_{i_g}, \tag{10.10}$$

In this setting, Assumption 3.9 is also relaxed to include the common factors:

ASSUMPTION 10.1. *x_{git} are generated from:*

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

10.2. Identification and Estimation

The Mean Cluster estimator presented in section 5 estimates consistently the parameters of interest of model 10.10 with the simple modification of inclusion of time and cluster-time dummies by exploiting the different moment conditions derived in this subsection.

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

$$\begin{aligned} y_{git} - y_{g.t} = \rho_g(y_{git-1} - y_{g.t-1}) + (x_{git} - x_{g.t})'\beta_g \\ + x'_{git}\lambda_{git} - x'_{g.t}\lambda_{g.t} + \varepsilon_{git} - \varepsilon_{g.t}. \end{aligned} \tag{10.11}$$

Where the cluster-time specific averages are equal to:

$$\frac{\sum_i y_{git}}{N_g} = \alpha_g + \gamma_t + \tau_{gt} + \rho_g \frac{\sum_i y_{git-1}}{N_g} + \frac{\sum_i x_{git}}{N_g}\beta_g + \frac{\sum_i x'_{git}\lambda_{git}}{N_g} + \frac{\sum_i \varepsilon_{git}}{N_g}. \tag{10.12}$$

We can just rename the transformed variables as:

$$\tilde{y}_{git} = \rho_g \tilde{y}_{git-1} + \tilde{x}'_{git}\beta_g + \tilde{x}'_{git\lambda_{git}} + \tilde{\varepsilon}_{git}. \tag{10.13}$$

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

$$\mathbb{E}(\tilde{u}_{git}\tilde{x}_{gis}) = 0 \quad \forall s \in (1, 2, ..., T), i \in (1, 2, ..., N_g), g \in (1, 2, ..., m), \tag{10.14}$$

$$\mathbb{E}(\tilde{u}_{git}\tilde{y}_{git-1}) = 0 \quad \forall t \in (1, 2, ..., T), i \in (1, 2, ..., N_g), g \in (1, 2, ..., m). \tag{10.15}$$

ASSUMPTION 10.2. *The matrix $\mathbb{E}(\tilde{z}_{git}\tilde{z}'_{git})$ is full rank.*

11. MONTE CARLO EXPERIMENT

In this section, I present a Monte Carlo simulation experiment performed to test the proposed estimators for the baseline model and the extensions of the baseline.

For this purpose, I generate 100 datasets from three different data generating processes

called DGP 1, DGP 2 and DGP 3. I use these datasets to test the proposed Mean Cluster estimators that are appropriate for DGP 1, the Bayesian estimator proposed for DGP 2 and the Mean Cluster estimator using the time demeaned variables for DGP 3.

In the following subsections I describe the Monte Carlo simulation design with more detail as well as the results.

11.1. The design

11.1.1. DGP 1 In order to test the estimation method proposed for model 1.1, I conduct a simulation experiment using a data generating processes that is similar to the DGP used by Hsiao et al. (1998). [9]

The main differences with the DGP of Hsiao et al. (1998) are: 1. inclusion of cluster effects instead of individual specific effects, 2. inclusion of multiplicative cluster-individual-time specific effects, 3. inclusion of correlated cluster specific effects, 4. the variance and variance-covariance are cluster specific and they are generated from Gamma and Wishart distributions. [10]

In particular, the DGP 1 is :

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

with $\rho_g = \bar{\rho} + \alpha_{2,g}$, $\bar{\rho} = 0.6$, $\beta_{git} = \bar{\beta} + \alpha_{3,g} + \lambda_{git}$ and $\bar{\beta} = \begin{pmatrix} 0.5 \\ 0.8 \end{pmatrix}$.

The cluster effects $\alpha_{j,g}$ are generated from a normal distribution centered at 0 and with variance equal to $\sigma^2_{\alpha,jg}$. The cluster specific variance $\sigma^2_{\alpha,jg}$ is sampled from a Gamma distribution with an inverse scale parameter equal to 1 and a shape parameter equal to 1.

The cluster effects added to the mean coefficient vector are generated from a multivariate normal distribution centered at 0 and with full variance-covariance matrix that is cluster specific. This variance-covariance matrix is sampled from a Whisart distribution with full scale matrix $V_{\alpha,3} = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$ and with 3 degrees of freedom.

The cluster-individual-time specific effects added to the mean coefficient vector are generated from a multivariate normal distribution centered at 0 with a variance covariance matrix that is cluster specific and sampled from a Wishart distribution. This Wishart distribution is parametrized by a full scale matrix $V_\lambda = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$ and 3 the degrees of freedom.

The disturbance term is generated from a normal distribution centered at 0 with a cluster heteroskedastic variance. The variance is sampled from a Gamma distribution inverse scale and shape parameters equal to 1.

The regressors x_{git} follow a stationary autoregressive processes similar to the process used by Hsiao et al. (1998). The key difference is that I allow for correlation with the cluster effects:

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

[9] I use modification of the DGP proposed by Hsiao et al. (1998) because the MG estimator is one of the most standard methods for estimation of dynamic heterogeneous panel data models.

[10] Detailed results of the simulation experiment are presented in Appendix D.

with $\mu_g \sim N(\iota_K, \sigma^2_{\mu,g} I_K)$, and $\sigma^2_{\mu,g} \sim Gamma(s_\mu, is_\mu)$ where $s_\mu = 1$ is the inverse scale parameter and $is_\mu = 1$ is the shape parameter.

The disturbance term of the regressors equation is sampled from the a normal distribution centered at 0 with variance that is cluster specific and generated from a Gamma distribution with scale parameter and shape parameters equal to 1.

11.1.2. DGP 2 In order to test the Bayesian estimator proposed for model 9.6, I modified DGP 1 by replacing the additive cluster specific effects with cluster-individual additive effects. The sample size is equal to 2 clusters, 50 individuals and 3 time periods.

11.1.3. DGP 3 In order to test the proposed estimator for the model presented in Section 10 I generate data from the following DGP with one common factor that is cluster specific and one common factor that affects the whole population:

$$y_{git} = \alpha_{1,g} + \rho_g y_{git-1} + x'_{git}\beta_{git} + \Lambda_g f_t + \Lambda_{gi} f_{gt} + \varepsilon_{git},$$

$$x_{git} = \mu_g(1-\phi) + \phi x_{git-1} + \alpha_{1,g} + \alpha_{3,g} + F_g f_t + F_{gi} f_{gt} + \omega_{git},$$

The factor loadings Λ_g, Λ_{gi}, F_g and F_{gi} are equal to 1. The common factors are stationary and generated as:

$f_t = 0.6 f_{t-1} + \epsilon_f$

$f_{gt} = 0.6 f_{gt-1} + \epsilon_{f_g}$

with $E[\epsilon_f] = 0$, $E[\epsilon_f^2] = 0.04$, $E[\epsilon_{f_g}] = 0$ and $E[\epsilon_{f_g}^2]$ is cluster specific and generated from a sigma distribution with scale and location parameters equal to 1.

11.2. The Results

11.2.1. DGP 1 In Table 1, I present the bias (relative bias*100 below the bias) and RMSE of the estimated mean parameters of interest. The estimates are obtained for 100 simulations for a sample with 10 groups, 100 individuals per group and time dimension equal to 3. The results are striking and demonstrate the power of the simple method presented.

Additionally, I performed the simulation experiment with different sample sizes varying the number of groups and the number of individuals per group. I present the results in Figures 1 to 8. In Figures 1 to 4, I plot the relative bias and the RMSE of the estimated mean parameters as a function of the number of individuals per cluster with the number of clusters equal to 2 and time observations equal to 3 and 6. In Figures 5 to 8, I plot the relative bias and the RMSE for the mean parameters s a function of the number of clusters and the number of individuals per cluster equal to 50 and time observations equal to 3 and 6.

The results show that when the data is balanced, the proposed Mean Cluster estimators have lower relative bias and RMSE than the Pooled OLS when the number of individuals within cluster grows or when the number of clusters grow. In particular, when the time dimension is equal to three the Mean Cluster OLS estimator has lower relative bias than OLS for all parameters. But the RMSE of the Mean Cluster OLS is close to the RMSE of OLS. This can be explained by the fact that in the Mean Cluster OLS we ignore the variance-covariance of the model.

When the data is weakly unbalanced, the proposed Mean Cluster estimators have lower

RMSE than the Pooled OLS if the time dimension is fixed and the number of individuals in the cluster grows. When the data is strongly unbalanced the best estimator is the Mean Cluster Pooled OLS. Morever, the Mean Cluster FGLS estimator does not perform correctly with strong unbalancedness. Finally, the Mean Cluster FGLS estimation does not perform well when time dimension is equal to 3 (The graphs of the simulation for unbalanced data for DGP 1 are presented in the Appendix D).

11.2.2. DGP 2 The simulation results for DGP 2 are presented in Table 2. They show that inclusion of the prior information of the initial observations produces estimates with low bias. The RMSE of the estimated autoregressive parameter is low but the RMSE of the slope coefficients is high. One reason for the high RMSE of the estimated slope coefficients could be the correlation between slope coefficients. These correlation impedes the convergence of the Hamiltonian Monte Carlo Algorithm. The reason is that the under the presence of parameter correlation, the HMC algorithm cannot explore efficiently all the posterior parameter space. A solution could be the implementation of a block Gibs sampler.

11.2.3. DGP 3 The results show that when the data is balanced, the proposed Mean Cluster estimators have lower RMSE than the Pooled OLS when the number of individuals within cluster grows or when the number of clusters grow (Graphs 13 to 12).

In all DGP, the Mean Cluster estimators seem unstable when the number of clusters is equal to 2. This is natural because the ratio of observed number of clusters is small with respect to the number of clusters. This happens because in the Monte Carlo experiment design I assume that the cluster specific parameters follow a continuous probability distribution. This means that one needs an increasing number of clusters to estimate correctly the mean cluster parameters. In this setting, it is clear that it not possible to learn anything from the global mean.

Table 1: Results Simulation DGP 1.

Pooled OLS		**Mean Cluster OLS**		**Mean Cluster FGLS**			
				$\tau = 0$		$\tau = 1$	
Bias	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**	**Bias**	**RMSE**
			ρ				
0.27	0.28	0.00	0.04	-0.01	0.04	0.00	0.06
45.71		0.55		-1.26		-0.46	
			β_1				
-0.22	0.35	0.11	0.35	-0.03	0.50	-0.02	0.57
-44.44		21.61		-5.66		-3.42	
			β_2				
-0.39	0.47	0.03	0.28	0.01	0.54	0.00	0.50
-49.20		3.64		0.65		0.33	

Note: τ represents the vaule of the regularization parameter used for estimation of the variance covariance components (See subsection 5.1).

Table 2: Results Simulation DGP 2

ρ			β_1			β_2		
Bias	**Bias%**	**RMSE**	**Bias**	**Bias%**	**RMSE**	**Bias**	**Bias%**	**RMSE**
0.05	7.83	0.07	0.03	6.08	0.34	0.03	4.11	0.33

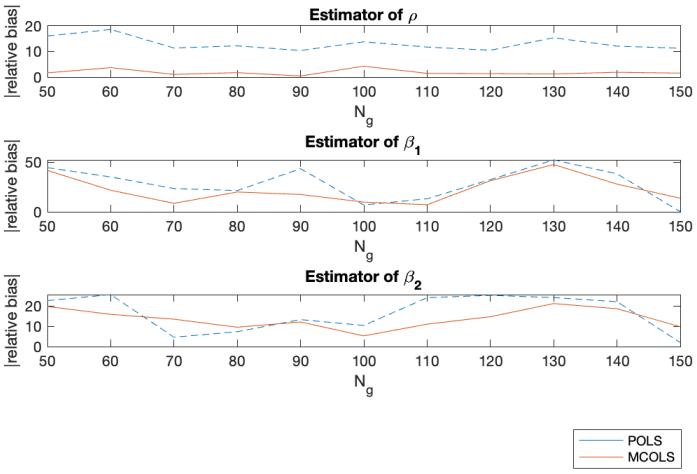


Figure 1: DGP 1: Relative Bias of estimated parameter as a function of the number of individuals per cluster with fixed $m = 2$, $T = 3$ (Balanced Panel).

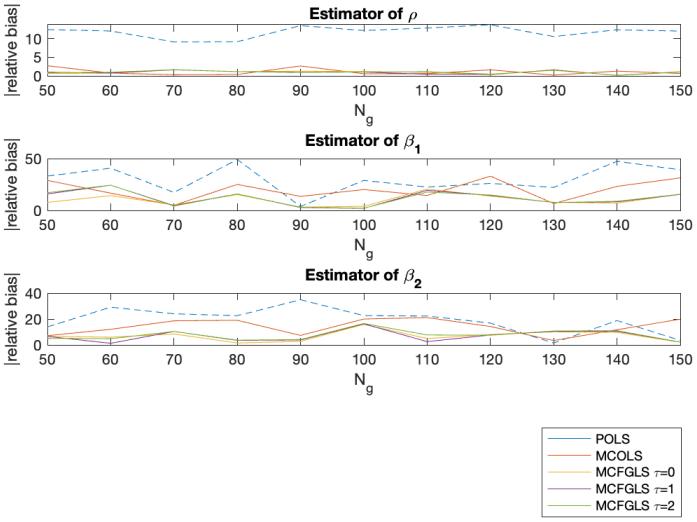


Figure 2: DGP 1: Relative Bias of estimated parameter as a function of the number of individuals per cluster with fixed $m = 2$, $T = 6$ (Balanced Panel).

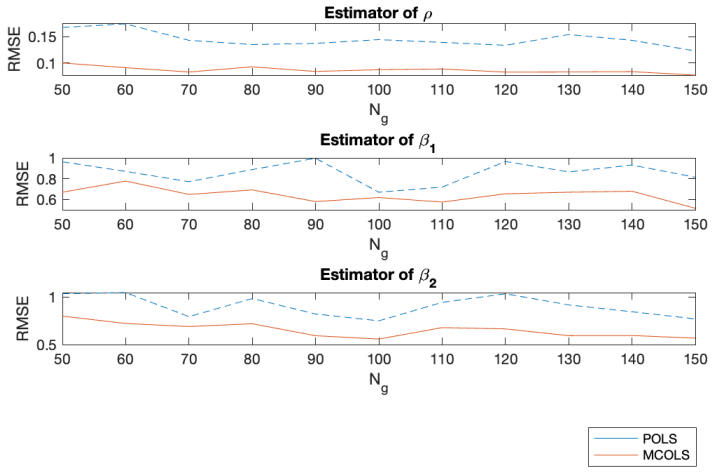


Figure 3: DGP 1: RMSE of estimated parameter as a function of the number of individuals per cluster with fixed $m = 2$, $T = 3$ (Balanced Panel).

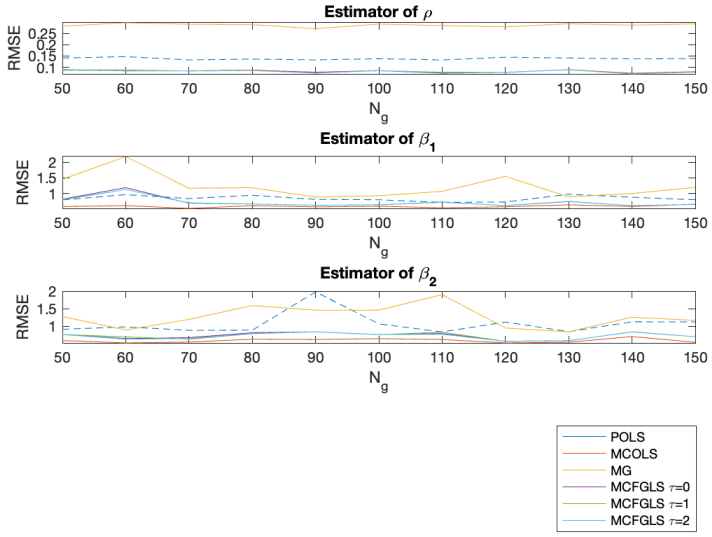


Figure 4: DGP 1: RMSE of estimated parameter as a function of the number of individuals per cluster with fixed $m = 2$, $T = 6$ (Balanced Panel).

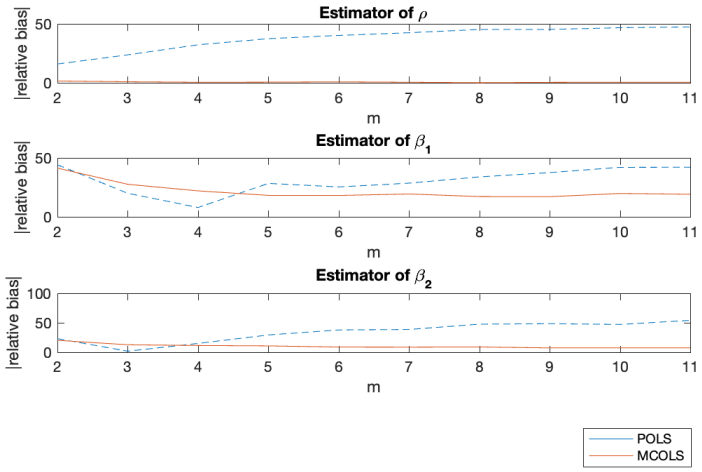


Figure 5: DGP 1: Relative Bias of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 3$ (Balanced Panel).

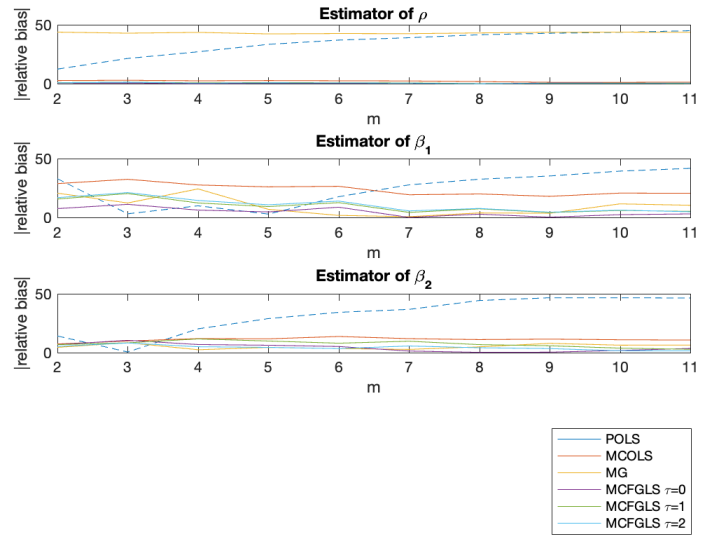


Figure 6: DGP 1: Relative Bias of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 6$ (Balanced Panel).

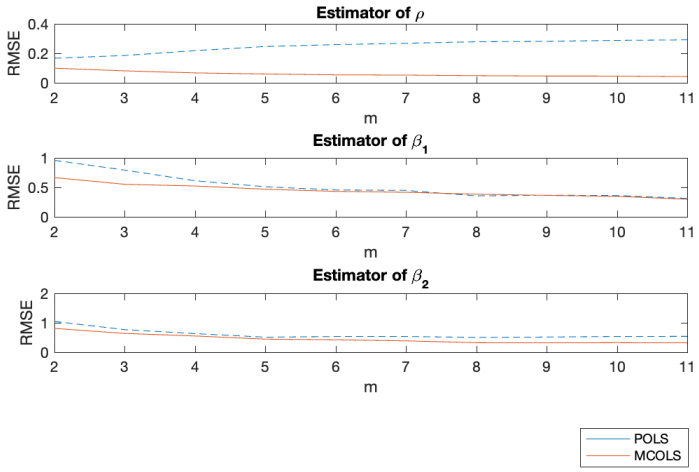


Figure 7: DGP 1: RMSE of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 3$ (Balanced Panel).

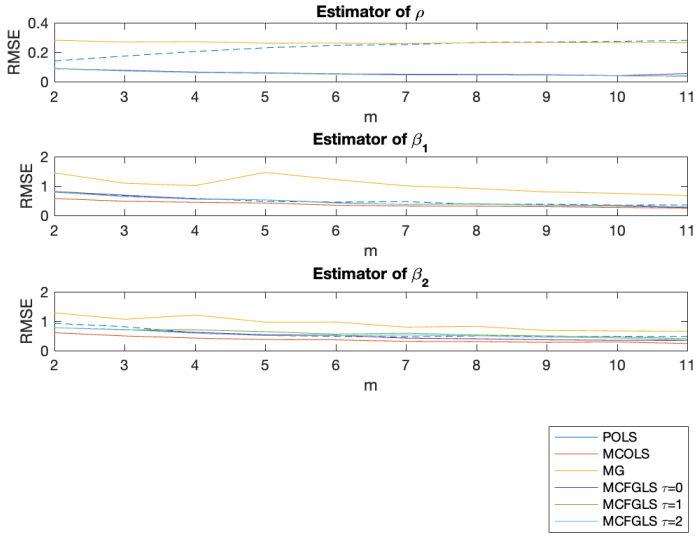


Figure 8: DGP 1: RMSE of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 6$ (Balanced Panel).

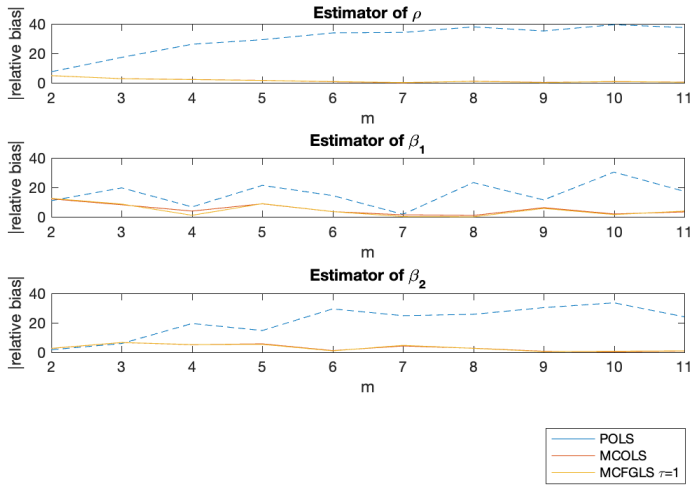


Figure 9: DGP 3: Relative Bias of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 3$ (Balanced Panel).

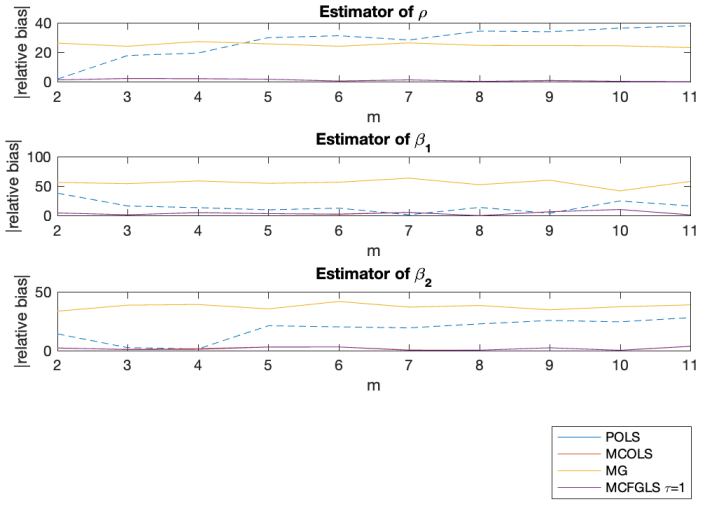


Figure 10: DGP 3: Relative Bias of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 6$ (Balanced Panel).

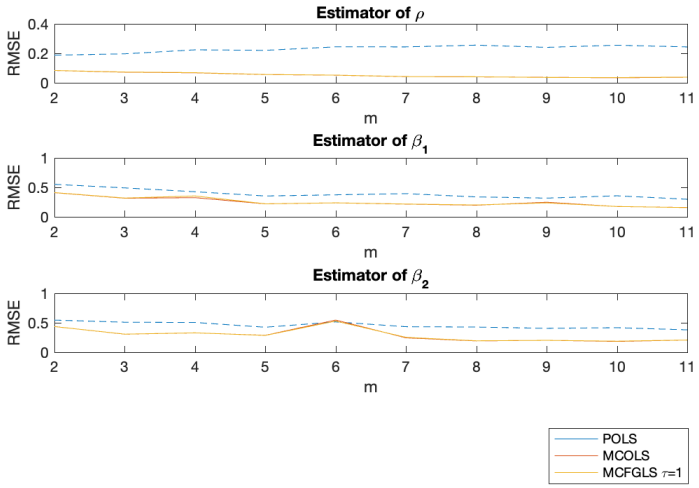


Figure 11: DGP 3: RMSE of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 3$ (Balanced Panel).

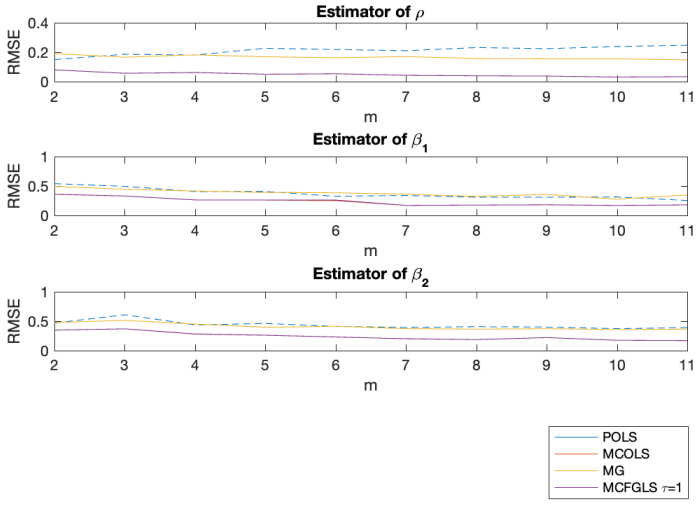


Figure 12: DGP 3: RMSE of estimated parameter as a function of the number clusters with fixed $N_g = 50$, $T = 6$ (Balanced Panel).

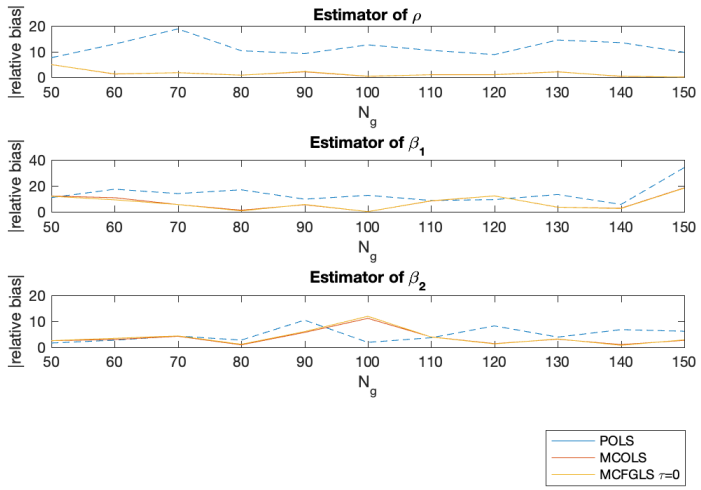


Figure 13: DGP 3: Relative bias of estimated parameter as a function of the number clusters with fixed $m = 2$, $T = 3$ (Balanced Panel).

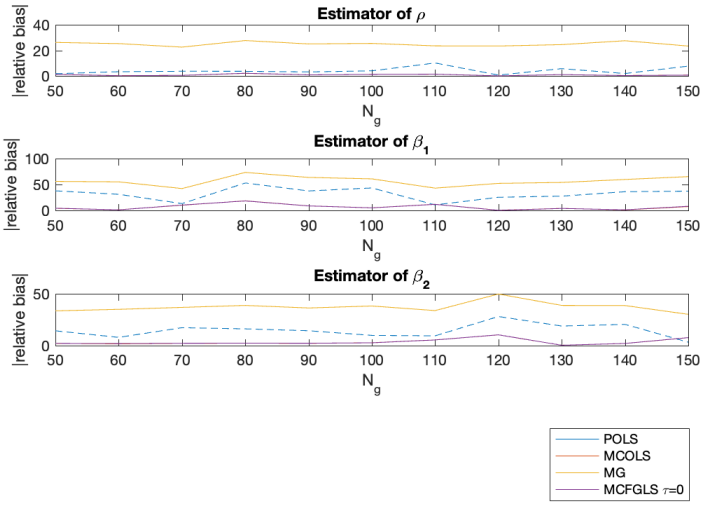


Figure 14: DGP 3: Relative bias of estimated parameter as a function of the number clusters with fixed $m = 2$, $T = 6$ (Balanced Panel).

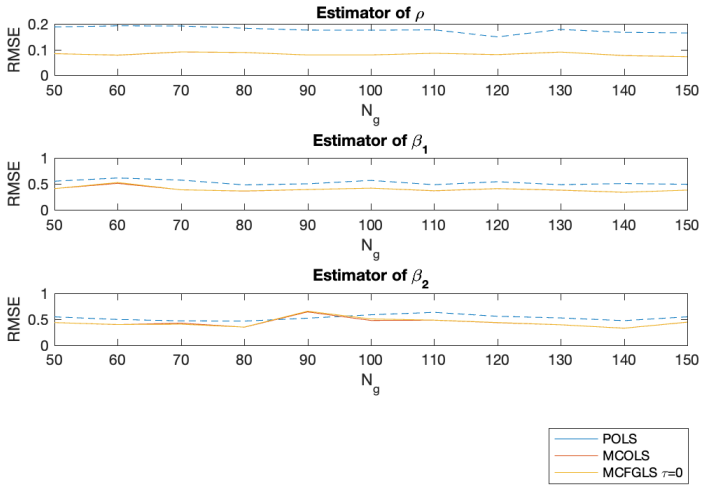


Figure 15: DGP 3: RMSE of estimated parameter as a function of the number clusters with fixed $m = 2$, $T = 3$ (Balanced Panel).

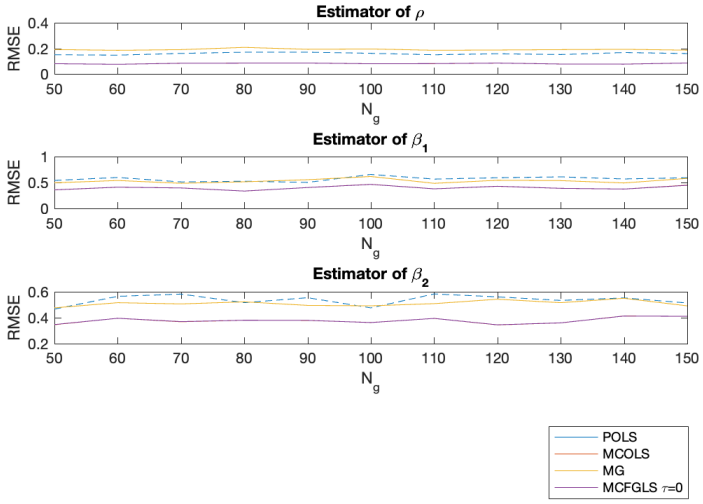


Figure 16: DGP 3: RMSE of estimated parameter as a function of the number clusters with fixed $m = 2$, $T = 6$ (Balanced Panel).

12. EMPIRICAL APPLICATION

In this section, I present an empirical example in order to illustrate the use of the methods presented. In particular, I use the Mean Cluster estimators to estimate the effect of enrollment of private school on the achievement of students.

12.1. Estimation of a value-added model of learning

In this example, I investigate the causal effect of private school on student achievement. Andrabi et al. (2011) studied this effect using a value-added model of learning and emphasized the importance of persistence. They used a data set obtained by means of a cluster sampling scheme. They surveyed 4031 students to obtain information on anthropometrics, and family characteristics in each period. More specifically, they consider the following value-added model for student i in period t:

$$score_{it} = \alpha_i + \rho score_{it-1} + \beta private_{it} + x'_{it}\theta + \epsilon_{it} \tag{12.1}$$

with:
α_i representing the individual specific effect,
$score_{it}$ is the score of mathematics for student i in period t,
$private_{it}$ is a dummy variable taking value 1 if student i is enrolled in a private school in period t,
x_{it} is a vector of control variables for student i in period t.

The authors faced three main empirical challenges: i) the time dimension is equal to 3, ii) the scores of Mathematics, Urdu, and English have measurement error causing autocorrelation of order 1 in the disturbance term and iii) the presence of individual specific unobserved heterogeneity. In order to eliminate the individual specific effects, the authors first-differenced the model. But first-differencing the model in combination with the autocorrelation of the disturbance term leaves no extra lags of the dependent variable to perform GMM estimation with instrumental variables. As a result, they use the score of other subjects as instrumental variables under the key assumption that the measurement error is uncorrelated across subjects. But a failure of this assumption invalidates their identification strategy.

In order to avoid making the assumption of non-correlation of the scores across subjects, I propose an alternative strategy that uses a model in levels with heterogeneous effects of private school on achievement across villages. This could be plausible if village characteristics have an effect on students achievements. For instance, it is possible that students in richer villages have higher effects. Additionally, I allow for the presence of common global factors as well as a common cluster factor in order to capture cross-sectional dependence of students within villages and across villages. More specifically, I propose the following extended value-added model for student i in cluster g at period t is given by:

$$score_{git} = \alpha_{gi} + \tau_{gt} + \gamma_t + \rho_g score_{git-1} + \beta_g private_{git} + x'_{git}\theta_g + \epsilon_{it} \tag{12.2}$$

where α_{gi} is a cluster-individual specific effect, τ_{gt} represents a common cluster effect, and γ_t represents a common global factor.

In this new model, I deal with the cluster-individual specific effects by following Mundlack approach by projecting them into the cluster-individual specific means of the regres-

sors such that $\alpha_{git} = \bar{z}'_{gi.}\pi_g + \omega_{gi}$ with $\bar{z}_{gi} = \frac{\sum_t z_{git}}{T}$ and z_{git} contains the control variables and private school. Additionally, I time demean the dependent variable and the regressors following the identification strategy presented in subsection 10.2 such that we end up with the model:

$$\begin{aligned} score_{git} - \bar{score}_{g.t} = \quad & \rho_g(score_{git-1} - score_{g.t-1}) + \beta_g(private_{git} - private_{g.t}) \\ & + (x_{git} - \bar{x}_{g.t})'\theta_g + (\bar{z}_{gi.} - \bar{z}_{g.})'\pi_g + \epsilon_{git} - \bar{\epsilon}_{g.t} + \omega_{gi} - \bar{\omega}_{g.}. \end{aligned} \tag{12.3}$$

In this transformed model, the endogeneous regressors are the time demeaned lag of the dependent variable and the time demeaned private school. Without first differencing the model, it is possible to identify the parameters of interest using the second lag of the time demeaned dependent variable and the time demeaned private school. Therefore, in the first stage I estimate the parameters of interest using 2SLS estimation and in the second stage I use the Mean Cluster estimator with a sample composed of the 112 villages clustered by district (I assign equal weights to each village). The estimated average persistence parameter is 0.65, and the average effect of private school on scores is 0.34 standard deviations. The estimated persistence parameter and private school effects using the Mean Cluster estimation are slightly larger than the ones obtained using pooled 2SLS controlling for the presence of individual effects with Mundlack approach (The estimated persistence parameter is equal to 0.66 and the estimated effect of the private school equal to 0.30 standard deviations). Using a Hausman type test comparing the pooled 2SLS with the Mean Cluster estimated parameter, I reject the null hypothesis in favor of the alternative one that there is village heterogeneity. Nevertheless, the results obtained from the test proposed must be taken with precaution since I do not investigate its statistical properties in this paper.

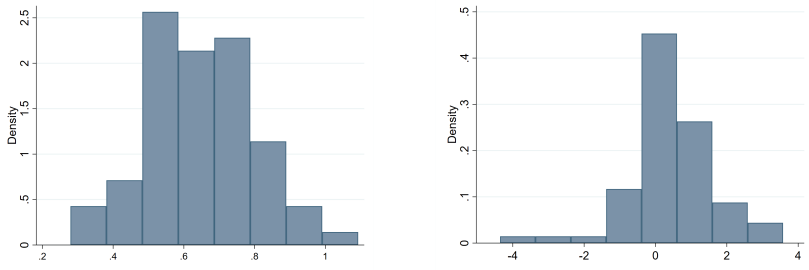
Additionally, I performed the estimation using villages as clusters. I use the 69 largest villages (villages with more than 100 students surveyed). In this case, the average estimated persistence parameter is equal to 0.63 and the average slope coefficient of private school is 0.40 standard deviations. The differences between district estimates and village estimates can be explained by the differences in the samples used for estimation.

The estimated persistence parameter using the mean cluster estimator is larger than the estimated value (0.12) presented by Andrabi et al. (2011). The estimated effect of private school on math score using Mean Cluster estimator is lower than the estimated value (0.40) of Andrabi et al. (2011). Additionally, the estimated effects using the Mean Cluster estimator are very similar to the estimated long run effects (0.31) using DID approach presented by Andrabi et al. (2011).

13. CONCLUSIONS

In this paper, I 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 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 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



(a) Persistent parameter (b) Private School Slope Coefficient

Figure 17: 2 Estimated Cluster Heterogenous Coefficients
Note: Estimated cluster specific parameters for 69 largest villages.

Smith (1995). When 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 estimates consistently the mean parameters.

As an extension of the baseline model, I consider a model with cluster-individual additive effects. In this setting, I suggest a hierarchical Bayesian estimator with a prior for the unknown initial conditions.

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 outperforms pooled OLS.

I can conclude from the simulation experiment, that the Mean Cluster estimators have lower RMSE than the MG 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.

ACKNOWLEDGEMENTS

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

REFERENCES

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017, November). When should you adjust standard errors for clustering? Working Paper 24003, National Bureau of Economic Research.

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.

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.

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.

Chamberlain, G. (1979). Analysis of covariance with qualitative data. Technical report, National Bureau of Economic Research.

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

Dhaene, G. and K. Jochmans (2015). Split-panel jackknife estimation of fixed-effect models. *The Review of Economic Studies 82*(3), 991–1030.

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.

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

H. Baltagi, B., S. Heun Song, and B. Cheol Jung (2001). The unbalanced nested error component regression model. *Journal of Econometrics 101*(2), 357 – 381.

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

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.

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. and M. Pesaran (2008). Random coefficient models. In L. Matyas and P. Sevestre (Eds.), *The Econometrics of Panel Data*, pp. 185–213. Springer.

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

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. Cham: Springer International Publishing.

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.

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

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.

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.

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

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.

14. APPENDIX A: PROOFS OF THEOREMS 1,2,3

14.1. Proof Theorem 1

Proof. It is known that:

$$\hat{\theta}_{g,GLS} = \theta_g + (\tilde{Z}_g'(\Omega_g^{-1})\tilde{Z}_g)^{-1}(\tilde{Z}_g'(\Omega_g^{-1})\tilde{w}_g),$$

with:

$$\tilde{w}_g = \tilde{X}_g diag(I_{T_{i_g}})\lambda_{gi} + \tilde{\epsilon}_g.$$

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

$$E(\hat{\theta}_{GLS,g} - \theta_g) = K_{g,n_g} + O_p(n_g^{-3/2}) = O_p(n_g^{-1})$$

I present the derivation of K_{g,n_g} in Appendix B.

Now, by assumption 3.2 we know that as N goes to infinity, $n_g \to \infty$ and $M = \sum_g n_g \to \infty$.

$$\sqrt{M}(\hat{\theta_g} - \theta_g) \sim N(0, Q_g),$$

where $Q_g = \underset{M\to\infty}{plim}(n_g^{-1}\tilde{Z}_g'(\Omega_g^{-1})\tilde{Z}_g)^{-1}$.
Thus, we can take advantage of the third dimension and still obtain an unbiased estimator even if T_{i_g} is fixed because N_g grows to infinity with the total number of individuals.

14.2. Proof Theorem 2

PROOF. As mentioned in section 5.1, the variance-covariance components stacked up in the vector η_g and estimated with a penalized LS approach as $(C_g'C_g + \tau I)^{-1}(C_g'\hat{R}_g)$ with C_g a full rank matrix obtained following the procedure proposed there. Now, for $\tau = 0$:

$$\sqrt{n_g}(\hat{\eta}_g - \eta_g) = (\frac{1}{n_g}\sum_j^{N_g}\sum_t^{T_{i_g}} C_{git}C_{git}')^{-1}(\frac{1}{n_g}\sum_j^{N}\sum_t^{T} C_{git}\nu_{git}).$$

Then, under assumption 13:

$$\sqrt{n_g}(\hat{\eta}_g - \eta_g) \xrightarrow{d} (0, \sigma^2_{Now,since}$$

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

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

14.3. Proof Theorem 3

PROOF. This follows from the property of sum of normal distributed vectors.

15. APPENDIX B: BIAS DERIVATION 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:

$$y_{gi} = \tilde{F}_g y_{gi0} + \tilde{C}_g X_{gi}\beta_g + \tilde{C}_g\tilde{X}_{gi}\lambda_{gi} + \varepsilon_{gi}, \tag{15.4}$$

where:

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

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

$$y_g = F_g y_{g0} + C_g x_g \beta_g + C_g\tilde{X}_g\lambda_g + \varepsilon_g, \tag{15.5}$$

where:
$F_g = diag(\tilde{F}_g)$,
$C_g = diag(\tilde{C}_g)$,
$\tilde{X}_g = diag(\tilde{X}_{gi})$,

Also, I know that the estimator per cluster is given by:

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

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:

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

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

Following **?**, I find that:

$$\begin{aligned} E[\hat{\theta}_g - \theta_g] = -(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1}[\bar{Z}_g'\Omega_g^{-1}C_g\bar{Z}_g(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1}e_1 + \\ e_1 tr((\bar{Z}_g'C_g\Omega_g^{-1}\bar{Z}_g)^{-1}\bar{Z}_g)(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1} + \\ e_1 e_1'(\bar{Z}_g'\Omega_g^{-1}\bar{Z}_g)^{-1}e_1 E[U_g'C_g'\Omega_g^{-1}C_g U_g U_g'C_g'\Omega_g^{-1}U_g)] + o(n_g^{-1}) \end{aligned} \tag{15.6}$$

This can be rewritten as:

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

16. APPENDIX C: CORRELATED CLUSTER EFFECTS FRAMEWORK

In this appendix, I present the model assumptions for a setting with cluster correlated effects. The identification and estimation strategy proposed in sections 4 and 5are still valid. Nevertheless, the asymptotic distribution of the Mean Cluster estimator is different as the one presented in section 6. Therefore, I also present the derivation of the asymptotic distribution of the mean cluster estimators.

16.1. Correlated cluster effects framework

ASSUMPTION 16.1. *The proportion of observed clusters q converges to 1.*

ASSUMPTION 16.2. *The number of clusters m is a monotonic function of* n_g, $\frac{\sqrt{m}}{n_g} \to 0$ *and* $m(n_g) \to \infty$.

ASSUMPTION 16.3. *Cluster-specific persistence parameter*

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

$$\rho_g = \bar{\rho} + \alpha_{1,g},$$

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

The cluster specific persistence parameter $\rho_g \in (-1, 1)$ is decomposed into two parts:

$E[\rho_g] = \bar{\rho}$ that is the mean persistence parameter and $\alpha_{1,g}$ that captures the heterogeneity across clusters.

ASSUMPTION 16.4. *Cluster-individual-time specific coefficients*

$$\beta_{git} = \bar{\beta} + \alpha_{2,g} + \lambda_{git},$$

$$E[\alpha_{2,g}\alpha'_{2,g'}] = \begin{cases} \Delta_{\alpha_{2,g}} & if \quad g = g' \\ 0 & otherwise. \end{cases},$$

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

The unobserved coefficient vector is given by $\beta_{git} \in \mathbb{R}^K$ and is equal to $\bar{\beta} + \alpha_{2,g} + \lambda_{git}$ where $\bar{\beta}$ is the mean coefficient vector, $\alpha_{2,g}$ captures the heterogeneity across clusters and λ_{git} captures the additional multiplicative heterogeneity over time for each individual of cluster g.

ASSUMPTION 16.5. *Correlated cluster specific effects with covariates.*

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

The cluster specific random components are conditionally dependent to the covariates allowing for correlated random coefficients at the cluster level.

ASSUMPTION 16.6. *No correlation of cluster-individual-time effects with the covariates.*

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

In this assumption, I state that the residual multiplicative heterogeneity is not correlated to the covariates. This implies that $E[\beta_{git} | x_{gi1}, x_{gi2}, ..., x_{giT}, y_{git-1}] = \beta_g$.

ASSUMPTION 16.7. *Non cross correlation between specific effects random components.*

$$(\alpha_{m,g}\lambda'_{git}) = 0 \quad for \quad m \in \{1, 2\} \quad and \quad \forall g, t, i.$$

In this assumption, I state that the group-individual-time specific effect is not cross correlated to the group heterogeneity.

16.2. Asymptotic behaviour of Mean Cluster Estimator

THEOREM 16.1. *Under assumptions presented in the previous subsection, $\hat{\bar{\theta}}$ is a consistent and normal distributed estimator.*

$$\sqrt{m}(\hat{\bar{\theta}} - \bar{\theta}) \sim N(0, \triangle_\alpha).$$

PROOF. It is known that:

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

$$\hat{\theta}_{g,GLS} = \bar{\theta} + \alpha_g + \xi_g,$$

with:

$$w_g = \tilde{X}_g diag(I_{T_{i_g}})\lambda_{gi} + \epsilon_g.$$

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

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

Thus, we can rewrite the mean-cluster GLS estimator as:

$$\hat{\bar{\theta}} = \bar{\theta} + \frac{1}{m}\sum_g^m \alpha_g + \frac{1}{m}\sum_g^m(\xi_g - \delta_{g,n_g}) + \frac{1}{m}\sum_g^m \delta_{g,n_g},$$

$$\sqrt{m}(\hat{\bar{\theta}} - \bar{\theta}) = \frac{1}{\sqrt{m}}\sum_g^m \alpha_g + \frac{1}{\sqrt{m}}\sum_g^m(\xi_g - \delta_{g,n_g}) + \frac{1}{\sqrt{m}}\sum_g^m \delta_{g,n_g},$$

$$\sqrt{m}(\hat{\bar{\theta}} - \bar{\theta}) = \frac{1}{\sqrt{m}}\sum_g^m \alpha_i + \frac{1}{\sqrt{m}}\sum_g^m(\xi_g - \delta_{g,n_g}) + \frac{\sqrt{m}}{n_g}(\frac{1}{m}\sum_g^m K_{g,n_g}) + O_p(\frac{\sqrt{m}}{n_g^{3/2}}).$$

Now, using assumption 12 we have that:

$$\sqrt{m}(\hat{\bar{\theta}} - \bar{\theta}) \sim N(0, \triangle_\alpha).$$

17. APPENDIX D: EXTENDED MONTE CARLO EXPERIMENT

17.1. The design

The sample sizes are $m \in \{2, 3, ..., 10\}$ with $N_g \in \{50, 51, ..., 150\}$ and $T_{i_g} \in \{3, 6\}$.

Additionally, the data is generated with two different patterns of random unbalancedness. The first one is weak and the second is strong (See table **??**). The unbalancedness level is measured by the following three measures that are an extension of the ones presented by H. Baltagi et al. (2001):

$$c_1 = m/\bar{N}\sum_g(1/N_g),$$

with $\bar{N} = \sum_g N_g/m$.

$$c_2 = N/\bar{T}\sum_g\sum_{i_g}(1/T_{i_g}),$$

with $N = \sum_g N_g$, $\bar{T} = \sum_g\sum_{i_g} T_{i_g}/N$.

$$c_3 = NT/\bar{NT} \sum_g \sum_{i_g} (1/N_g T_{i_g}).$$

with $\bar{NT} = \sum_g \sum_{i_g} (1/N_g T_{i_g})/N$.

These measures take value 1 when the three level panel data is fully balanced and they decrease with higher levels of unbalancedness. The unbalancedness patterns are presented in Table 3. For instance, in row weak-a the sample has 2 clusters with 26 and 12 individuals. The cluster with 26 individuals has 9 individuals with 3 time observations and 17 individuals with 6 time observations.

Table 3: Unbalancedness patterns and sample sizes

Unbalanced Pattern	c1	c2	c3	m	N_g	T_{i_g}
None	1	1	1	2	50	3
None	1	1	1	2	50	6
Strong	0.4756	1	0.6560	2	1(50), 1(8)	3
Weak	0.8733	.77	0.6737	2	1(40), 1(19)	34(6), 16(2)
Weak	0.8733	1	0.8876	2	1(40), 1(19)	3
Strong	0.4756	0.7509	0.5187	2	1(50), 1(8)	12(5), 12(2),26(6)

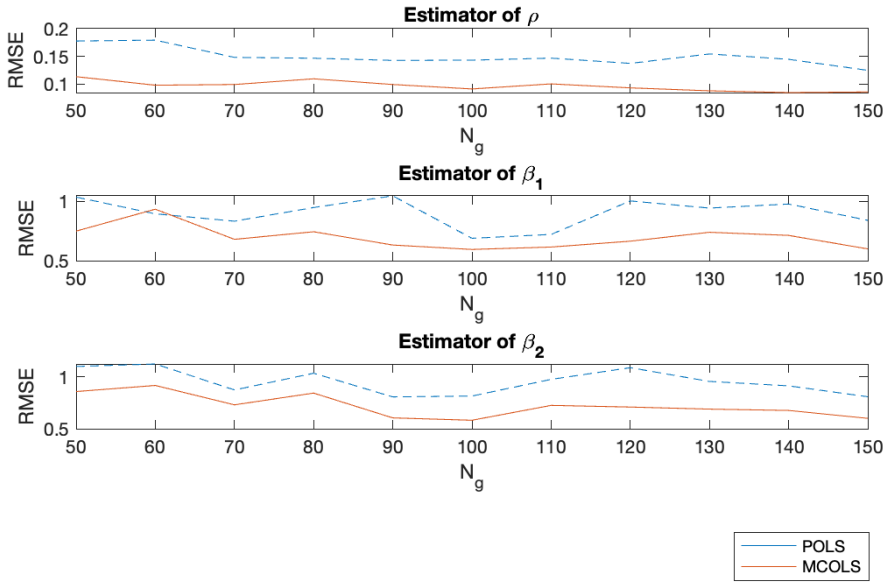


Figure 18: DGP 1: RMSE of estimated parameters as a function of the number individuals per cluster with fixed $m = 2$, $T = 3$ (Weak unbalanced Panel).

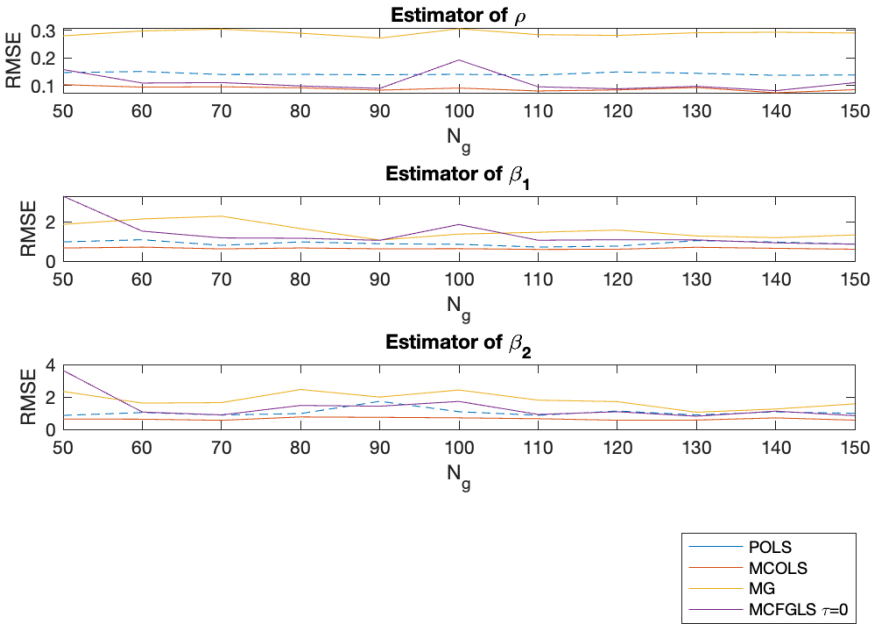


Figure 19: DGP 1: RMSE of estimated parameters as a function of the number individuals per cluster with fixed $m = 2$, $T = 6$ (Weak unbalanced Panel).

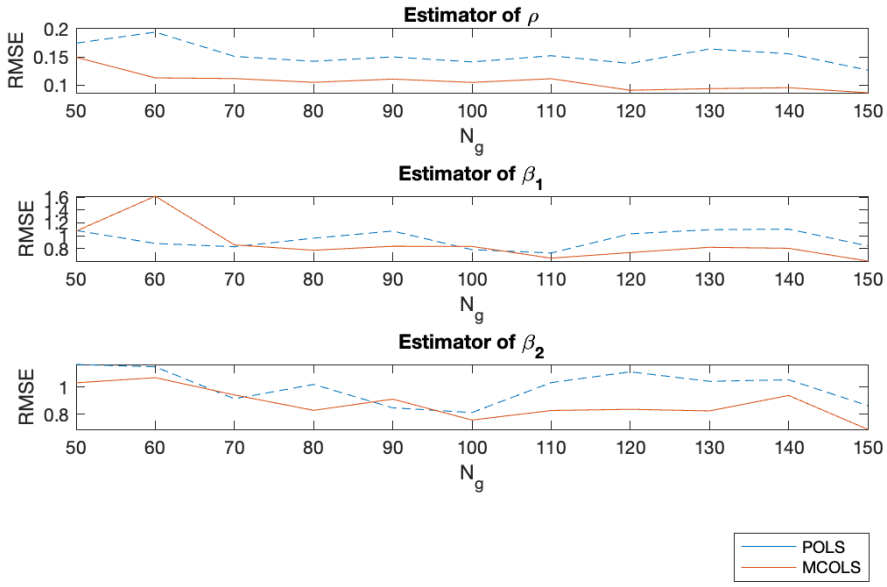


Figure 20: DGP 1: RMSE of estimated parameters as a function of the number individuals per cluster with fixed $m = 2$, $T = 3$ (Strong unbalanced Panel).

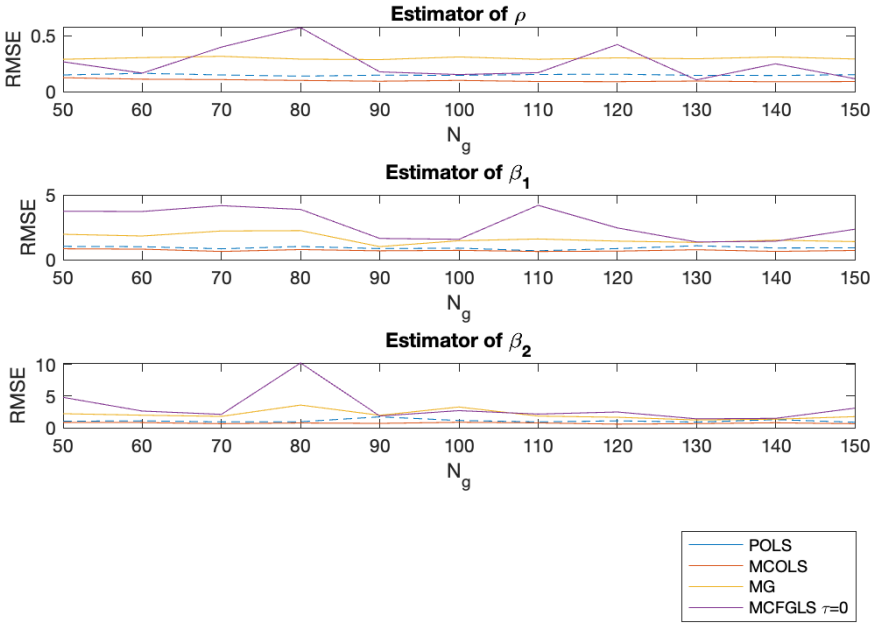


Figure 21: DGP 1: RMSE of estimated parameters as a function of the number individuals per cluster with fixed $m = 2$, $T = 6$ (Strong unbalanced Panel).

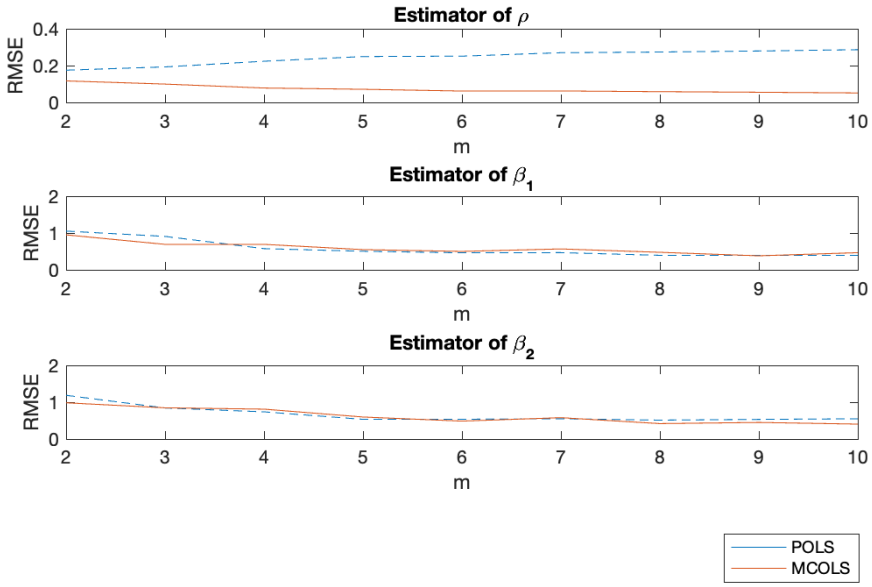


Figure 22: DGP 1: RMSE of estimated parameters as a function of the number of cluster with fixed $N_g = 50$, $T = 3$ (Weak unbalanced Panel).

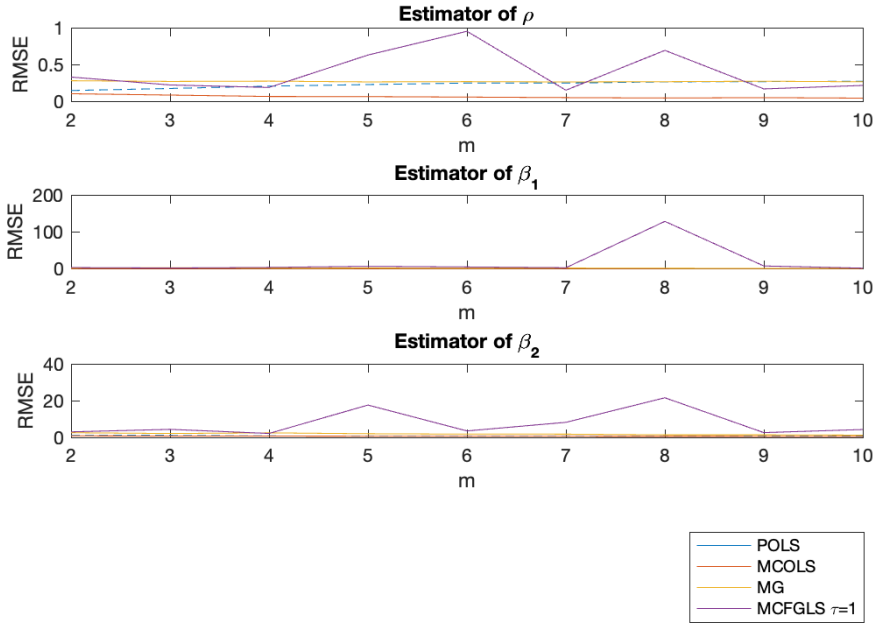


Figure 23: DGP 1: RMSE of estimated parameters as a function of the number of clusters with fixed $N_g = 50$, $T = 6$ (Weak unbalanced Panel).

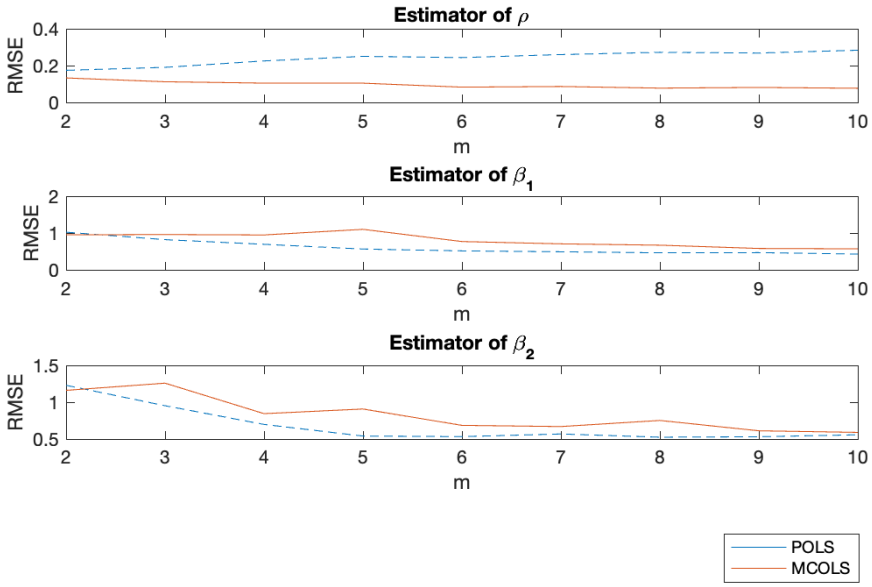


Figure 24: DGP 1: RMSE of estimated parameters as a function of the number of clusters with fixed $N_g = 50$, $T = 3$ (Strong unbalanced Panel).

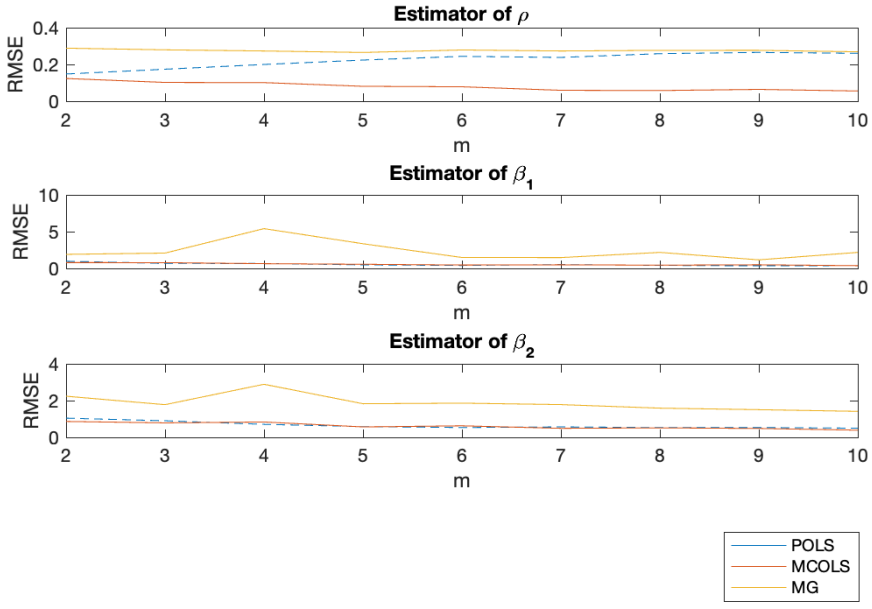


Figure 25: DGP 1: RMSE of estimated parameters as a function of the number of clusters with fixed $N_g = 50$, $T = 6$ (Strong unbalanced Panel).