Grouped Patterns of Heterogeneity in Panel Data[*]

Stéphane Bonhomme
University of Chicago

Elena Manresa
MIT Sloan

Final version: December 2014

Abstract

This paper introduces time-varying grouped patterns of heterogeneity in linear panel data models. A distinctive feature of our approach is that group membership is left unrestricted. We estimate the parameters of the model using a "grouped fixed-effects" estimator that minimizes a least-squares criterion with respect to all possible groupings of the cross-sectional units. Recent advances in the clustering literature allow for fast and efficient computation. We provide conditions under which our estimator is consistent as both dimensions of the panel tend to infinity, and we develop inference methods. Finally, we allow for grouped patterns of unobserved heterogeneity in the study of the link between income and democracy across countries.

JEL codes: C23.

Keywords: Discrete heterogeneity, panel data, fixed effects, clustering, democracy.

[*] We thank the co-editor Elie Tamer and two anonymous referees for their help in improving the paper. We also thank Daron Acemoglu, Daniel Aloise, Dante Amengual, Manuel Arellano, Jushan Bai, Alan Bester, Fabio Canova, Victor Chernozhukov, Denis Chetverikov, David Dorn, Kirill Evdokimov, Iván Fernández-Val, Lars Hansen, Jim Heckman, Han Hong, Bo Honoré, Andrea Ichino, Jacques Mairesse, Mónica Martínez-Bravo, Serena Ng, Taisuke Otsu, Eleonora Patacchini, Peter Phillips, Jean-Marc Robin, Enrique Sentana, Martin Weidner, and seminar participants at various venues for useful comments. Support from the European Research Council/ ERC grant agreement n^{0}263107 and from the Ministerio de Economía y Competitividad Grant ECO2011 − 26342 is gratefully acknowledged. All errors are our own.

1 Introduction

There is ample evidence that workers, firms or countries differ in many dimensions that are unobservable to the econometrician. In practice, applied researchers face a trade-off between using flexible approaches to model unobserved heterogeneity, and building parsimonious specifications that are well adapted to the data at hand. The goal of this paper is to propose a flexible yet parsimonious approach to allow for unobserved heterogeneity in a panel data context.

A common approach is to model heterogeneity as unit-specific, time-invariant fixed-effects. Fixed-effects approaches are attractive as they allow for unrestricted correlation between unobserved effects and covariates. However, in models with as many parameters as individual units, estimates of common parameters are subject to an "incidental parameter" bias that may be substantial in short panels (Nickel, 1981), and the fixed-effects themselves are often poorly estimated. In addition, standard fixed-effects approaches are arguably restrictive as they assume that unobserved heterogeneity is constant over time.

This paper proposes a framework that allows for clustered time patterns of unobserved heterogeneity that are common within groups of individuals. The group-specific time patterns and individual group membership are left unrestricted, and are estimated from the data. In particular, as in fixed-effects, our time-varying specification allows for general forms of covariates endogeneity. The main assumption is that the number of distinct individual time patterns of unobserved heterogeneity is relatively small.

A simple linear model with grouped patterns of heterogeneity takes the following form:

$$y_{it} = x_{it}'\theta + \alpha_{g_i t} + v_{it}, \quad i = 1, ..., N, \quad t = 1, ..., T, \tag{1}$$

where the covariates x_{it} are contemporaneously uncorrelated with v_{it}, but may be arbitrarily correlated with the group-specific unobservables $\alpha_{g_i t}$. The group membership variables $g_i \in \{1, ..., G\}$ and the group-specific time effects α_{gt}, for $g \in \{1, .., G\}$, are unrestricted. Units in the same group share the same time profile α_{gt} (for example, all i such that $g_i = 1$ share the profile α_{1t}). The number of groups G is to be set or estimated by the researcher. Beyond model (1) we study in detail two extensions: a model with additive time-invariant fixed-effects η_i in addition to the time-varying grouped effects $\alpha_{g_i t}$, and a model with group-specific coefficients θ_{g_i}.

Potential applications of model (1) and its extensions include social interaction models for panel data where group-level interactions are subsumed in $\alpha_{g_i t}$ (Blume *et al.*, 2010), or tests of full risk-sharing in village economies (Townsend, 1994). Unlike most applications of social interactions and risk sharing, our approach allows to estimate the reference groups, under the assumption that group membership remains constant over time. In a different perspective, grouped patterns of heterogeneity can be useful to model interdependence across individual units over time. Compared with existing spatial dependence models for panel data (e.g., Sarafidis and Wansbeek, 2012), model (1) allows the researcher to estimate the spatial weights matrix.

Our estimator, which we will refer to as "grouped fixed-effects" (GFE), is based on an optimal grouping of the N cross-sectional units, according to a least-squares criterion. Units whose time profiles of outcomes– net of the effect of covariates– are most similar are grouped together in estimation. In the absence of covariates in model (1), the estimation problem coincides with the standard minimum sum-of-squares partitioning problem, and a simple computational method is given by the "*kmeans*" algorithm (Forgy, 1965, Steinley, 2006). We take advantage of recent advances in the clustering literature to build fast and reliable computational routines.[1]

We derive the statistical properties of the grouped fixed-effects estimator in an asymptotic where N and T tend to infinity simultaneously. In our framework, N can grow substantially faster than T, in contrast with models with unit-specific fixed-effects. While fixed-effects estimators generally suffer from a O(1/T) bias as N/T tends to a constant (Arellano and Hahn, 2007), we show that the GFE estimator is consistent and asymptotically normal as N/T^{ν} tends to zero for some $\nu > 0$, provided groups are well separated and errors v_{it} satisfy suitable tail and dependence conditions. This property, which has also been noted in models with time-invariant discrete heterogeneity (Hahn and Moon, 2010), is a consequence of group classification improving very fast as the number of time periods increases. In particular, our results provide a formal justification for clustering methods.[2]

As the two dimensions of the panel diverge, the GFE estimator is asymptotically equivalent to the infeasible least squares estimator with known population groups. As a consequence, in a large-T perspective standard errors are unaffected by the fact that group membership has been estimated. In short panels, however, group misclassification may contribute to the finite-sample dispersion of the estimator. For this reason, we also study the properties of the GFE estimator for fixed T as N tends to infinity. In a Monte Carlo exercise calibrated to the empirical application, we provide evidence that using the GFE estimator in combination with an estimate of its fixed-T variance yields reliable inference for the population parameters.

We use our approach to study the effect of income on democracy in a panel of countries that spans the last part of the twentieth century. In an influential paper, Acemoglu *et al.* (2008) find that the positive association between income and democracy disappears when controlling for additive country- and time-effects. They interpret the country fixed-effects as reflecting long-run, historical factors that have shaped the political and economic development of countries.

In the context of this application, the grouped fixed-effects model allows for time-varying unobservables in a period that is characterized by a large number of transitions to democracy, and it is well suited to deal with the short length of the panel ($T = 7$). Grouped patterns are also consistent with

[1] Stata codes, which make use of a Fortran executable program, are available as supplementary material.

[2] Related results in statistics include Pollard (1981, 1982) on minimum sum-of-squares partitioning, and Bryant and Williamson (1978) in a class of likelihood models. To our knowledge, ours is the first paper to establish conditions under which the clustering estimator in *kmeans* becomes consistent as N and T tend to infinity.

the empirical observation that regime types and transitions tend to cluster in time and space (e.g., Gleditsch and Ward, 2006, Ahlquist and Wibbels, 2012). An early conceptual framework is laid out in Huntington (1991)'s work on the "third wave of democracy", which argues that international and regional factors– such as the influence of the Catholic Church or the European Union– may have induced grouped patterns of democratization. We find robust evidence of heterogeneous, group-specific paths of democratization in the data.

Related literature and outline. Our modelling of grouped heterogeneity is related to, but different from, finite mixture models. These models rely on assumptions that restrict the relationship between unobserved heterogeneity and covariates.[3] In contrast, and in close analogy with fixed-effects, our approach leaves that relationship unspecified. In fact, the group membership variables g_i may be viewed as indexing the N time-varying paths of unit-specific unobserved heterogeneity. The key assumption is that at most G of these paths are distinct from each other. This imposes a restriction on the *support* of unobserved heterogeneity, while leaving other features of the relationship with observables unrestricted.[4]

The grouped fixed-effects model is also related to factor-analytic, "interactive fixed-effects" models (Bai, 2009). Indeed, model (1) has a factor-analytic structure, as: $\alpha_{g_i t} = \sum_{g=1}^{G} \mathbf{1}\{g_i = g\}\alpha_{gt}$. We take advantage of this mathematical connection to establish consistency of the GFE estimator, and to study a class of information criteria to select the number of groups. Our theoretical and numerical results suggest that, in relatively short panels and when the data have a grouped structure, the parsimony of the GFE estimator may provide a useful alternative to interactive fixed-effects.

Finally, this paper is not the first one to rely on grouped structures for modelling unobserved heterogeneity in panels. Bester and Hansen (2013) show that grouping individual fixed-effects can result in gains in precision, in a setup where the grouping of the data is known. Lin and Ng (2012) consider a random coefficients model and use the time-series regression estimates to classify individual units into several groups. None of these two papers allows for time-varying unobserved heterogeneity.[5]

The outline of the paper is as follows. We introduce the grouped fixed-effects estimator in Section 2. We derive its asymptotic properties in Section 3. We use the GFE approach to study the relationship between income and democracy in Section 4, and conclude in Section 5. Additional material may be found in a supplementary appendix.

[3] See the monographs by McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for recent advances in this area.

[4] In this sense, our approach is reminiscent of sparsity assumptions in the literature on high-dimensional modelling (e.g., Tibshirani, 1996).

[5] Group models and clustering approaches have also been used to search for "convergence clubs" in the empirical growth literature; see for example Canova (2004), and Phillips and Sul (2007). See also Sun (2005). Similar techniques have been proposed in the statistical analysis of network data (e.g., Bickel and Chen, 2009, Choi *et al.*, 2012).

2 The grouped fixed-effects estimator

In the first part of this section we introduce the grouped fixed-effects (GFE) estimator in several models. In the second part we provide computational methods.

2.1 Models and estimators

Model (1) contains three types of parameters: the parameter vector $\theta \in \Theta$, which is common across individual units; the group-specific time effects $\alpha_{gt} \in \mathcal{A}$, for all $g \in \{1, ..., G\}$ and all $t \in \{1, ..., T\}$; and the group membership variables g_i, for all $i \in \{1, ..., N\}$, which map individual units into groups. The parameter spaces Θ and $\mathcal{A}$ are subsets of $\mathbb{R}^K$ and $\mathbb{R}$, respectively. We denote as α the set of all α_{gt}'s, and as γ the set of all g_i's. Thus, $\gamma \in \Gamma_G$ denotes a particular grouping (i.e., partition) of the N units, where Γ_G is the set of all groupings of $\{1, ..., N\}$ into at most G groups.

The covariates vector x_{it} may include strictly exogenous regressors and lagged outcomes. The model also allows for time-invariant regressors under certain support conditions. Moreover, x_{it} and $\alpha_{g_i t}$ are allowed to be arbitrarily correlated. We will state precise conditions in the next section.

The grouped fixed-effects estimator in model (1) is defined as the solution of the following minimization problem:

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2, \tag{2}$$

where the minimum is taken over all possible groupings $\gamma = \{g_1, ..., g_N\}$ of the N units into G groups, common parameters θ, and group-specific time effects α.

For given values of θ and α, the optimal group assignment for each individual unit is:

$$\widehat{g}_i\left(\theta, \alpha\right) = \underset{g\in\{1,...,G\}}{\operatorname{argmin}} \sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{gt}\right)^2, \tag{3}$$

where we take the minimum g in case of a non-unique solution. The GFE estimator of (θ, α) in (2) can then be written as:

$$\left(\widehat{\theta}, \widehat{\alpha}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2, \tag{4}$$

where $\widehat{g}_i\left(\theta, \alpha\right)$ is given by (3). The GFE estimate of g_i is then simply $\widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right)$.

Unlike standard finite mixture modelling, which specifies the group probabilities as parametric or semiparametric functions of observed covariates (e.g., McLachlan and Peel, 2000), grouped fixed-effects leaves group membership unrestricted. In the supplementary appendix we show that the GFE estimator maximizes the pseudo-likelihood of a mixture-of-normals model, where the mixing probabilities are unrestricted and individual-specific. In this perspective, the grouped fixed-effects approach may be seen as a point of contact between finite mixtures and fixed-effects.

Extension 1: unit-specific heterogeneity. The GFE framework can be combined with additive time-invariant fixed-effects:

$$y_{it} = x_{it}'\theta + \alpha_{g_i t} + \eta_i + v_{it}, \tag{5}$$

where η_i are N unrestricted parameters. Letting $\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} w_{it}$, the following equation in deviations to the mean:

$$y_{it} - \overline{y}_i = (x_{it} - \overline{x}_i)'\,\theta + \alpha_{g_i t} - \overline{\alpha}_{g_i} + v_{it} - \overline{v}_i, \tag{6}$$

has the same structure as model (1), and can be estimated using grouped fixed-effects.

Extension 2: heterogeneous coefficients. Another extension is to allow for group-specific effects of covariates:

$$y_{it} = x_{it}'\theta_{g_i} + \alpha_{g_i t} + v_{it}, \tag{7}$$

and define the following GFE estimator:

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta^G\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta_{g_i} - \alpha_{g_i t}\right)^2, \tag{8}$$

where here θ contains all θ_g's.

Other extensions of the baseline model are possible. For example, one could allow for unit-specific heterogeneity and heterogeneous coefficients at the same time. In addition, in the supplementary appendix we show how to incorporate prior information on the groups or the group-specific time effects when such information is available.

Nonlinear models. Grouped patterns of heterogeneity may be introduced in nonlinear models as well. A general M-estimator formulation based on a data-dependent function $m_{it}(\cdot)$ is as follows:

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta^G\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} m_{it}\left(\theta, \alpha_t, g_i\right). \tag{9}$$

This framework covers likelihood models as special cases.[6] In particular, it encompasses static and dynamic discrete choice models. However, studying the statistical properties of GFE in nonlinear models exceeds the scope of this paper.

2.2 Computation

One can see, from (4), that the grouped fixed-effects estimator minimizes a piecewise-quadratic function, where the partition of the parameter space is defined by the different values of $\widehat{g}_i\left(\theta, \alpha\right)$, for

[6] A GFE estimator in a likelihood setup is obtained by taking $m_{it}\left(\theta, \alpha_t, g_i\right) = -\ln f\left(y_{it}|x_{it}; \theta, \alpha_t, g_i\right)$, where $f(\cdot)$ denotes a parametric density function.

$i = 1, ..., N$. However, the number of partitions of N units into G groups increases steeply with N, making exhaustive search virtually impossible.

The following algorithm uses a simple iterative strategy to minimize (4).

Algorithm 1 *(iterative)*

1. *Let* $\left(\theta^{(0)}, \alpha^{(0)}\right) \in \Theta \times \mathcal{A}^{GT}$ *be some starting value. Set* $s = 0$.

2. *Compute for all* $i \in \{1, ..., N\}$*:*

$$g_i^{(s+1)} = \underset{g \in \{1,...,G\}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta^{(s)} - \alpha_{gt}^{(s)} \right)^2. \tag{10}$$

3. *Compute:*

$$\left(\theta^{(s+1)}, \alpha^{(s+1)}\right) = \underset{(\theta,\alpha) \in \Theta \times \mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{t=1}^{N} \left(y_{it} - x_{it}'\theta - \alpha_{g_i^{(s+1)}t} \right)^2. \tag{11}$$

4. *Set* $s = s + 1$ *and go to Step 2 (until numerical convergence).*

Algorithm 1 alternates between two steps. In the "assignment" step, each individual unit i is assigned to the group g_i whose vector of time effects is closest (in an Euclidean sense) to her vector of residuals $y_{it} - x_{it}'\theta$. In the "update" step, θ and α are computed using an OLS regression that controls for interactions of group indicators and time dummies.[7] The objective function is non-increasing in the number of iterations, and numerical convergence is typically very fast. However, the solution depends on the chosen starting values. Drawing starting values at random and selecting the solution that yields the lowest objective provides a practical approach in low-dimensional problems.

For larger-scale problems, we take advantage of the close connection between (4) and the well-studied *kmeans* clustering algorithm (Forgy, 1965), and develop a more efficient computational routine, Algorithm 2, which exploits recent advances in data clustering (Hansen *et al.*, 2010). In the supplementary appendix we compare the performance of the two algorithms against exact computational methods. While Algorithm 1 recovers the global minimum in a small-scale dataset for G = 3 groups, we provide evidence that Algorithm 2 also reaches the global minimum when G = 10. Computer codes that allow to compute the GFE estimator in generic panels are available as online material. Even though computation on large datasets is currently challenging, recently developed heuristic and exact methods– some of which are surveyed in the supplementary appendix– suggest the potential for vast improvements in speed and accuracy.

[7] As written, the solution of the algorithm may have empty groups. A simple modification consists in re-assigning one individual unit to every empty group, as in Hansen and Mladenović (2001). Note that doing so automatically decreases the objective function.

3 Asymptotic properties

In this section we characterize the asymptotic properties of the grouped fixed-effects estimator as N and T tend to infinity in model (1). Extensions of the main theorems for models (5) and (7) can be found in the supplementary appendix.

3.1 The setup

Consider the following data generating process:

$$y_{it} = x_{it}'\theta^0 + \alpha^0_{g_i^0 t} + v_{it}, \tag{12}$$

where $g_i^0 \in \{1, ..., G\}$ denotes group membership, and where the ^{0} superscripts refer to true parameter values. We assume for now that the number of groups $G = G^0$ is known, and we defer the discussion on estimation of the number of groups until the end of this section.

Let $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$ be the infeasible version of the GFE estimator where group membership g_i, instead of being estimated, is fixed to its population counterpart g_i^0:

$$\left(\widetilde{\theta}, \widetilde{\alpha}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{g_i^0 t}\right)^2. \tag{13}$$

This is the least-squares estimator in the pooled regression of y_{it} on x_{it} and the interactions of population group dummies and time dummies.

The main result of this section provides conditions under which estimated groups converge to their population counterparts, and the GFE estimator defined in (2) is asymptotically equivalent to the infeasible least-squares estimator $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$, when N and T tend to infinity and $N/T^\nu \to 0$ for some $\nu > 0$. In particular, this allows T to grow considerably more slowly than N (when $\nu \gg 1$). Before discussing the general case of model (12), we provide an intuition in a simple case.

Intuition in a simple case. Consider a simplified version of model (12) in which group-specific effects are time-invariant, $\theta^0 = 0$ is known (no covariates), v_{it} are i.i.d. normal $(0, \sigma^2)$, and $G = G^0 = 2$:

$$y_{it} = \alpha^0_{g_i^0} + v_{it}, \quad g_i^0 \in \{1, 2\}, \quad v_{it} \sim iid\mathcal{N}(0, \sigma^2). \tag{14}$$

We assume that $\alpha_1^0 < \alpha_2^0$. The properties of GFE are different when group separation fails (e.g., when $\alpha_1^0 = \alpha_2^0$), as we discuss below.

In finite samples, there is a non-zero probability that estimated and population group membership do not coincide. Specifically, it follows from (3) that the probability of misclassifying into group 2 an individual who belongs to group 1 is:

$$\Pr\left(\widehat{g}_i\left(\alpha^0\right) = 2 \middle|\, g_i^0 = 1\right) = \Pr\left(\sum_{t=1}^{T}\left(\alpha_1^0 + v_{it} - \alpha_2^0\right)^2 < \sum_{t=1}^{T}\left(\alpha_1^0 + v_{it} - \alpha_1^0\right)^2\right) = \Pr\left(\overline{v}_i > \frac{\alpha_2^0 - \alpha_1^0}{2}\right).$$

That is:

$$\Pr\left(\widehat{g}_i\left(\alpha^0\right) = 2 \middle| \, g_i^0 = 1\right) = 1 - \Phi\left(\sqrt{T}\left(\frac{\alpha_2^0 - \alpha_1^0}{2\sigma}\right)\right), \tag{15}$$

where Φ denotes the standard normal cdf.

For fixed T, $\widehat{g}_i\left(\alpha^0\right)$ is inconsistent as N tends to infinity, because only the ith observation is informative about g_i^0. As a result, $\widehat{\alpha}$ generally suffers from an incidental parameter bias and is inconsistent. Nevertheless, (15) implies that the group misclassification probability tends to zero at an *exponential* rate, which intuitively means that the incidental parameter problem vanishes very rapidly as T increases.

Extending the analysis of model (14) to a more general setup raises two main challenges. First, consistency is not straightforward to establish since, as N and T tend to infinity, both the number of group membership variables g_i and the number of group-specific time effects α_{gt} tend to infinity, causing an incidental parameter problem in both dimensions.[8] Second, the argument leading to the exponential rate of convergence in (15) relies on i.i.d. normal errors. In order to bound tail probabilities under more general conditions, approximations based on a central limit theorem are not sufficient.

3.2 Consistency

Consider the following assumptions.

Assumption 1 *There exists a constant $M > 0$ such that:*

a. Θ and $\mathcal{A}$ are compact subsets of $\mathbb{R}^K$ and $\mathbb{R}$, respectively.

b. $\mathbb{E}\left(\|x_{it}\|^2\right) \leq M$, where $\|\cdot\|$ denotes the Euclidean norm.

c. $\mathbb{E}\left(v_{it}\right) = 0$, and $\mathbb{E}\left(v_{it}^4\right) \leq M$.

d. $\left|\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T \mathbb{E}\left(v_{it}v_{is}x_{it}'x_{is}\right)\right| \leq M$.

e. $\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N\left|\frac{1}{T}\sum_{t=1}^T \mathbb{E}\left(v_{it}v_{jt}\right)\right| \leq M$.

f. $\left|\frac{1}{N^2T}\sum_{i=1}^N\sum_{j=1}^N\sum_{t=1}^T\sum_{s=1}^T \mathrm{Cov}\left(v_{it}v_{jt}, v_{is}v_{js}\right)\right| \leq M$.

g. Let $\overline{x}_{g\wedge\widetilde{g},t}$ denote the mean of x_{it} in the intersection of groups $g_i^0 = g$, and $g_i = \widetilde{g}$.[9] For all groupings

[8] Note that the class of models considered in a recent paper by Hahn and Moon (2010) only covers *time-invariant* discrete unobserved heterogeneity. So their results do not apply here.

[9] Formally: $\overline{x}_{g\wedge\widetilde{g},t} = \frac{\sum_{i=1}^N \mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_i=\widetilde{g}\}x_{it}}{\sum_{i=1}^N \mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_i=\widetilde{g}\}}$. Note that $\overline{x}_{g\wedge\widetilde{g},t}$ depends on the grouping $\gamma = \{g_1, ..., g_N\}$, although we leave that dependence implicit for conciseness. In fact, Theorem 1 below remains true if, in Assumption 1.g, the average $\overline{x}_{g_i^0\wedge g_i,t}$ is replaced by the linear projection of x_{it} on the group indicators $\mathbf{1}\{g_i^0 = 1\}$, ..., $\mathbf{1}\{g_i^0 = G\}$, $\mathbf{1}\{g_i = 1\}$, ..., $\mathbf{1}\{g_i = G\}$, all of them interacted with time dummies.

$\gamma = \{g_1, ..., g_N\} \in \Gamma_G$ we define $\widehat{\rho}(\gamma)$ as the minimum eigenvalue of the following matrix:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x}_{g_i^0 \wedge g_i,t}\right)\left(x_{it} - \overline{x}_{g_i^0 \wedge g_i,t}\right)'.$$

Then $\operatorname{plim}_{N,T\to\infty} \min_{\gamma\in\Gamma_G} \widehat{\rho}(\gamma) = \rho > 0$.

In Assumption 1.a we require the parameter spaces to be compact. It is possible to relax this assumption and alternatively assume that the group-specific time effects α_{gt}^0 have finite moments, as in Bai (2009). However, allowing the group effects to follow non-stationary processes would require a different analysis, which is not considered in this paper. Similarly, we rule out non-stationary covariates and errors in Assumptions 1.b and 1.c, respectively.

Weak dependence conditions are required in Assumptions 1.d to 1.f. These are related to assumptions commonly made in the literature on large factor models (Stock and Watson, 2002, Bai and Ng, 2002). Assumption 1.d allows for lagged outcomes and general predetermined regressors, for example when $\mathbb{E}\left(v_{it}|x_{it}, x_{i,t-1}, ..., v_{i,t-1}, v_{i,t-2}, ...\right) = 0$. Assumptions 1.d and 1.f impose conditions on the time-series dependence of errors (and covariates), while Assumption 1.e restricts the amount of cross-sectional dependence. Note that the latter condition is satisfied in the special case where v_{it} are independent across units.

Lastly, Assumption 1.g is a relevance condition, reminiscent of full rank conditions in standard regression models. We require that x_{it} shows sufficient within-group variation over time and across individuals.[10] As a special case, the condition will be satisfied if x_{it} are discrete and, for all g, the conditional distribution of $(x_{i1}, ..., x_{iT})$ given $g_i^0 = g$ has strictly more than G points of support. As another special case, it can be shown that Assumption 1.g holds when x_{it} are i.i.d. normal.[11] Note also that Assumption 1.g allows for time-invariant regressors, provided that their support is rich enough.

We have the following result, where for conciseness we denote $\widehat{g}_i = \widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right)$ the GFE estimates of g_i^0, for all i.

Theorem 1 *(consistency) Let Assumption 1 hold. Then, as N and T tend to infinity:*

$$\widehat{\theta} \xrightarrow{p} \theta^0, \ \textit{and}$$

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha_{g_i^0 t}^0\right)^2 \xrightarrow{p} 0.$$

Proof. See Appendix A. ■

[10] Assumption 1.g is interestingly related to Assumption A in Bai (2009).

[11] To see this, let us suppose that $x_{it} \sim \mathcal{N}(0,1)$ for simplicity. Then $\max_{\gamma\in\Gamma_G}\sum_{i=1}^{N}\sum_{t=1}^{T}\overline{x}_{g_i^0\wedge g_i,t}^2$ is the maximum of $|\Gamma_G| \leq G^N$ random variables drawn from a χ_{DT}^2 distribution, where $D \leq G^2$, so that Assumption 1.g is satisfied.

3.3 Asymptotic distribution

Consider the following additional assumptions.

Assumption 2

a. For all $g \in \{1, ..., G\}$: $\operatorname{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} = \pi_g > 0$.

b. For all $(g, \widetilde{g}) \in \{1, ..., G\}^2$ such that $g \neq \widetilde{g}$: $\operatorname{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2 = c_{g,\widetilde{g}} > 0$.

c. There exist constants $a > 0$ and $d_1 > 0$ and a sequence $\alpha[t] \leq e^{-at^{d_1}}$ such that, for all $i \in \{1, ..., N\}$ and $g \in \{1, ..., G\}$, $\{v_{it}\}_t$ and $\{\alpha_{gt}^0\}_t$ are strongly mixing processes with mixing coefficients $\alpha[t]$.[12] *Moreover, $\mathbb{E}\left(\alpha_{gt}^0 v_{it}\right) = 0$ for all $g \in \{1, ..., G\}$.*

d. There exist constants $b > 0$ and $d_2 > 0$ such that $\Pr\left(|v_{it}| > m\right) \leq e^{1-\left(\frac{m}{b}\right)^{d_2}}$ *for all i, t, and $m > 0$.*

e. There exists a constant $M^* > 0$ such that, as N, T tend to infinity:

$$\sup_{i\in\{1,...,N\}} \Pr\left(\frac{1}{T}\sum_{t=1}^{T} \|x_{it}\| \geq M^*\right) = o\left(T^{-\delta}\right) \text{ for all } \delta > 0.$$

In contrast with consistency, we restrict the analysis of the asymptotic distribution to the case where the G population groups have a large number of observations and are well-separated (Assumptions 2.a and 2.b). The main asymptotic equivalence result does not hold uniformly with respect to the group-specific parameters. An example when group separation fails is when the number of groups in the population is strictly smaller than the number of groups postulated by the researcher (i.e., when $G^0 < G$). At the end of this section and in the supplementary appendix we come back to this important issue.

In Assumptions 2.c and 2.d we restrict the dependence and tail properties of v_{it}, respectively. Specifically, we assume that v_{it} are strongly mixing with a faster-than-polynomial decay rate (which strengthens the assumptions made in Assumption 1 regarding time-series dependence), with tails also decaying at a faster-than-polynomial rate. The process α_{gt}^0 is assumed to be strongly mixing, and to be contemporaneously uncorrelated with v_{it}. These conditions allow us to rely on exponential inequalities for dependent processes (e.g., Rio, 2000) in order to bound misclassification probabilities.[13]

Finally, in Assumption 2.e we impose a condition on the distribution of covariates x_{it}. This condition holds if covariates have bounded support or, alternatively, if they satisfy dependence and

[12] Note that $\alpha[t]$ is a conventional notation for strong mixing coefficients. We use this notation here, in the hope that this does not generate confusion with the group-specific time effects α_{gt}.

[13] It is possible to relax Assumptions 2.c-2.d and assume that v_{it} and α_{gt}^0 are strongly mixing with a polynomial decay rate, and that the marginal distribution of v_{it} has polynomial tails, i.e. that $\alpha[t] \leq at^{-d_1}$, and $\Pr\left(|v_{it}| > m\right) \leq m^{-d_2}$ for some constants $a \geq 1$, $d_1 > 1$, and $d_2 > 2$. It may then be shown that $\widehat{\theta} - \widetilde{\theta} = o_p\left(T^{-q}\right)$, provided that $\frac{(d_1+1)d_2}{d_1+d_2} > 4q + 1$.

tail conditions similar to the ones on v_{it}. However, strong mixing conditions may not necessarily hold when lagged outcomes (e.g., $y_{i,t-1}$) are included in the set of covariates. For example, Andrews (1984) discusses simple autoregressive models that are not strongly mixing. We show in Appendix B that Assumption 2.e is also satisfied when, in addition to strongly mixing covariates, the model allows for a lagged outcome with autoregressive coefficient $|\rho^0| < 1$, and the distribution of the initial conditions y_{i0} has thinner-than-polynomial tails.

The next result shows that the GFE estimator and the infeasible least squares estimator with known population groups (see equation (13)) are asymptotically equivalent under Assumptions 1 and 2. Note that, because of invariance to re-labelling of the groups, the results for group membership and group-specific effects are understood to hold given a suitable choice of the labels (see the proof for details).

Theorem 2 *(asymptotic equivalence) Let Assumptions 1 and 2 hold. Then, for all $\delta > 0$ and as N and T tend to infinity:*

$$\Pr\left(\sup_{i\in\{1,...,N\}} \left|\widehat{g}_i - g_i^0\right| > 0\right) = o(1) + o\left(NT^{-\delta}\right), \tag{16}$$

and:

$$\widehat{\theta} = \widetilde{\theta} + o_p\left(T^{-\delta}\right), \quad \text{and} \tag{17}$$

$$\widehat{\alpha}_{gt} = \widetilde{\alpha}_{gt} + o_p\left(T^{-\delta}\right) \quad \text{for all } g, t. \tag{18}$$

Proof. See Appendix B. ■

The following assumptions allow to simply characterize the asymptotic distribution of the least-squares estimator $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$. We denote as $\overline{x}_{gt}$ the mean of x_{it} in group $g_i^0 = g$.

Assumption 3

a. *For all i, j and t: $\mathbb{E}\left(x_{jt}v_{it}\right) = 0$.*

b. *There exist positive definite matrices Σ_θ and Ω_θ such that:*

$$\Sigma_\theta = \operatorname*{plim}_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x}_{g_i^0 t}\right)\left(x_{it} - \overline{x}_{g_i^0 t}\right)'$$

$$\Omega_\theta = \lim_{N,T\to\infty} \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathbb{E}\left[v_{it}v_{js}\left(x_{it} - \overline{x}_{g_i^0 t}\right)\left(x_{js} - \overline{x}_{g_j^0 s}\right)'\right].$$

c. *As N and T tend to infinity: $\frac{1}{\sqrt{NT}}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x}_{g_i^0 t}\right)v_{it} \xrightarrow{d} \mathcal{N}\left(0, \Omega_\theta\right)$.*

d. *For all (g, t): $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}\left(\mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_j^0 = g\}v_{it}v_{jt}\right) = \omega_{gt} > 0$.*

e. For all (g,t)*, and as* N *and* T *tend to infinity:* $\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathbf{1}\{g_i^0=g\}v_{it}\xrightarrow{d}\mathcal{N}\left(0,\omega_{gt}\right)$*.*

Assumptions 3.a-3.c imply that the least-squares estimator $\widetilde{\theta}$ has a standard asymptotic distribution. Assumption 3.a is satisfied if x_{it} are strictly exogenous or predetermined and observations are independent across units. As a special case, lagged outcomes may thus be included in x_{it} (although the assumption does not allow for *spatial* lags such as $y_{i-1,t}$). Similarly, Assumptions 3.d-3.e ensure that $\widetilde{\alpha}_{gt}$ has a standard asymptotic distribution.

The following result is a direct consequence of Theorem 2.

Corollary 1 *(asymptotic distribution) Let Assumptions 1, 2, and 3 hold, and let* N *and* T *tend to infinity such that, for some* $\nu>0$*,* $N/T^{\nu}\to 0$*. Then we have:*

$$\sqrt{NT}\left(\widehat{\theta}-\theta^0\right)\xrightarrow{d}\mathcal{N}\left(0,\Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1}\right),\tag{19}$$

and, for all (g,t)*:*

$$\sqrt{N}\left(\widehat{\alpha}_{gt}-\alpha_{gt}^0\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\omega_{gt}}{\pi_g^2}\right),\tag{20}$$

where π_g *is defined in Assumption 2, and where* Σ_θ*,* Ω_θ*, and* ω_{gt} *are defined in Assumption 3.*

Proof. See the supplementary appendix. ■

Under the conditions of Corollary 1, the GFE estimator of θ^0 is root-NT consistent and asymptotically normal in an asymptotic where T can increase polynomially more slowly than N. The GFE estimates of group-specific time effects are root-N consistent and asymptotically normal under the same conditions. Moreover, the estimated group membership indicators are uniformly consistent for the population ones as $N/T^{\nu}\to 0$ for some $\nu>0$, in the sense that: $\Pr\left(\sup_{i\in\{1,\dots,N\}}\left|\widehat{g}_i-g_i^0\right|>0\right)\to 0$. As a result:[14]

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widehat{g}_i t}-\alpha_{g_i^0 t}^0\right)^2=O_p\left(\frac{1}{N}\right).\tag{21}$$

These properties contrast with those of estimators that allow for unit-specific fixed-effects in combination with time fixed-effects. Given the interactive structure of model (12), "interactive fixed-effects" estimators are particularly relevant in our context. The interactive fixed-effects estimator of θ^0, as fixed-effects estimators in other settings, has a $O(1/T)$ bias in general when $N/T\to c>0$, see Theorem 3 in Bai (2009). In addition, the conditions for root-N consistency of the time-varying factors require that $N/T^2\to 0$, see Theorem 1 in Bai (2003).[15] Lastly, when using interactive fixed-effects the

[14] Equation (21) holds if: $\frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\mathbb{E}\left(\mathbf{1}\{g_i^0=g\}\mathbf{1}\{g_j^0=g\}v_{it}v_{jt}\right)=O(1)$, in addition to the conditions of Corollary 1. See the supplementary appendix for a proof.

[15] Theorem 1 in Bai and Ng (2002) does not rely on this condition, but yields a rate of $\min(\sqrt{N},\sqrt{T})$.

components $\alpha^0_{g_i^0 t}$ are estimated at a rate of $\min(\sqrt{N},\sqrt{T})$, see Theorem 3 in Bai (2003). These properties suggest that, when a grouped structure is a reasonable assumption, GFE may be better suited than interactive fixed-effects in panels of moderate length. Simulations calibrated to the empirical application, summarized below, are in line with this theoretical discussion.

3.4 Additional properties and extensions

Here we briefly discuss additional theoretical and numerical properties of the grouped fixed-effects estimator in models (1), (5), and (7). Details are provided in the supplementary appendix.

Inference. The large-N,T asymptotic analysis above provides conditions under which group membership estimation does not affect inference. In the supplementary appendix we discuss various estimators of the matrices defined in Assumption 3 that allow to conduct feasible inference under those conditions.

When T is kept fixed as N tends to infinity, in contrast, estimation of group membership matters for inference. In the supplementary appendix we extend previous results by Pollard (1981, 1982) to allow for covariates, and derive an analytical formula for the fixed-T variance of the GFE estimator. In this alternative asymptotic framework, the variance reflects the additional contribution of observations that are at the margin between two groups, so that an infinitesimal change in parameter values may entail re-classifying these observations.

A fixed-T asymptotic analysis is not directly informative to perform valid inference for the population parameters since, for fixed T, the GFE estimator $\left(\widehat{\theta},\widehat{\alpha}\right)$ is root-N consistent and asymptotically normal for a *pseudo-true* value $\left(\overline{\theta},\overline{\alpha}\right)$. This pseudo-true value, which minimizes an expected within-group sum of squared residuals, does not coincide with the true parameter value in general, but the difference between the two vanishes as T increases. A practical possibility to account for the effect of group membership estimation on inference is to use the GFE estimator in combination with a fixed-T consistent estimator of its variance. In the supplementary appendix we propose two such estimators: an estimator of the analytical variance formula, and a bootstrap-based estimator.

Choice of the number of groups. Following Bai and Ng (2002), we study in the supplementary appendix how to estimate the number of groups G^0 using information criteria. In addition, to explore the impact of misspecifying the number of groups, we analytically study a simple model with time-invariant group-specific effects, where the true number of groups is $G^0 = 1$ but the researcher postulates $G = 2$ (so $\alpha^0_1 = \alpha^0_2$). In this example, common parameter estimates are consistent for fixed T, but group-specific effects suffer from large biases. Moreover, specifying $G < G^0$ generally leads to biases on common parameters and group-specific effects. The choice of G, and the related issue of how inference on the model's parameters is affected by this choice, are difficult questions that deserve further investigation.

Simulation evidence. In order to assess the finite sample performance of the GFE estimator we conduct several exercises on simulated data. The designs mimic the cross-country dataset that we use in the empirical application ($N = 90$, $T = 7$). We find small probabilities of group misclassification (less than 10% when $G = 3$ and $G = 5$), and moderate biases on common parameters. Moreover, when comparing the GFE estimator to the interactive fixed-effects estimator on a simulated dataset with grouped heterogeneity, we find that the latter has large biases and imprecisely estimated components of unobserved heterogeneity. Finally, we compare different inference methods, and conclude that estimators of the fixed-T variance lead to more reliable inference for the population parameters. Details and additional exercises can be found in the supplementary appendix.

Extension 1: unit-specific heterogeneity. An equivalence result analogous to Theorem 2 holds in model (5), with additive time-invariant fixed-effects in addition to the time-varying grouped effects. The conditions given in the supplementary appendix allow for strictly exogenous covariates and lagged outcomes. One difference with the baseline analysis is that Assumption 1.g then involves deviations of covariates with respect to their unit-specific means, reflecting the fact that time-variation in x_{it} is necessary when a fixed-effect is included in the model. A second difference is the group separation condition: we require $\text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0 - \overline{\alpha}_g^0 - \alpha_{\widetilde{g}t}^0 + \overline{\alpha}_{\widetilde{g}}^0\right)^2 > 0$, where $\overline{\alpha}_g^0 = \frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}^0$. In the presence of additive fixed-effects, consistent estimation of group membership is only possible if the group-specific profiles are not parallel.

In model (5), GFE estimates of group membership indicators are consistent, and the equivalence result holds relative to an infeasible fixed-effects estimator. When covariates x_{it} are strictly exogenous, a result analogous to Corollary 1 holds. However, when x_{it} include a lagged outcome $y_{i,t-1}$, the fixed-effects estimator $\widetilde{\theta}$ suffers from a $O(1/T)$ bias (as in Nickel, 1981). Once group membership indicators have been consistently estimated using GFE, we suggest estimating θ using an instrumental variables strategy. We provide details on this two-step approach in the supplementary appendix.

Extension 2: heterogeneous coefficients. We also provide an asymptotic characterization of the GFE estimator in model (7) with group-specific coefficients. One difference with the baseline case is that group separation requires: $\text{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'\left(\theta_g^0 - \theta_{\widetilde{g}}^0\right) + \alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2 > 0$. Establishing separation conditions for interesting classes of nonlinear models, and developing methods to test these conditions or perform inference that is robust to lack of group separation, are important questions for future work.

4 Application: income and (waves of) democracy

The statistical association between income and democracy is an important stylized fact in political science and economics (Lipset, 1959, Barro, 1999). In an influential paper, Acemoglu, Johnson, Robinson and Yared (2008) emphasize the importance of accounting for factors that simultaneously affect economic and political development. Using panel data, they document that the positive effect of income on democracy disappears when including country fixed-effects in the regression. They argue that these results are consistent with countries having embarked on divergent paths of economic and political development at certain points in history, or *critical junctures*. Some of the examples they mention are the end of feudalism, the industrialization age, or the process of colonization. In this perspective, the fixed-effects are meant to capture these highly persistent historical events.

In this section, we revisit the evidence using the grouped fixed-effects approach in a regression of democracy (measured by the Freedom House indicator) on lagged democracy and lagged log-GDP per capita with unrestricted group-specific time patterns of heterogeneity α_{g_it}:[16]

$$democracy_{it} = \theta_1 democracy_{it-1} + \theta_2 logGDPpc_{it-1} + \alpha_{g_it} + v_{it}. \tag{22}$$

In the supplementary appendix we report the results of a number of alternative specifications.

Coefficient estimates: income and lagged democracy. Figure 1 shows the point-estimates and standard errors of income and lagged democracy coefficients for different values of the number of groups G, on the 1970-2000 balanced subsample of Acemoglu *et al.* (2008).[17] The right panel shows that the implied cumulative income effect $\theta_2/(1-\theta_1)$ sharply decreases from .25 in OLS to .10 for $G = 5$, and remains almost constant as G increases further. The left and middle panels show that this pattern is mostly driven by a decrease in the coefficient of lagged democracy. This is consistent with unobserved country heterogeneity being positively correlated with lagged democracy, causing an upward bias in OLS.

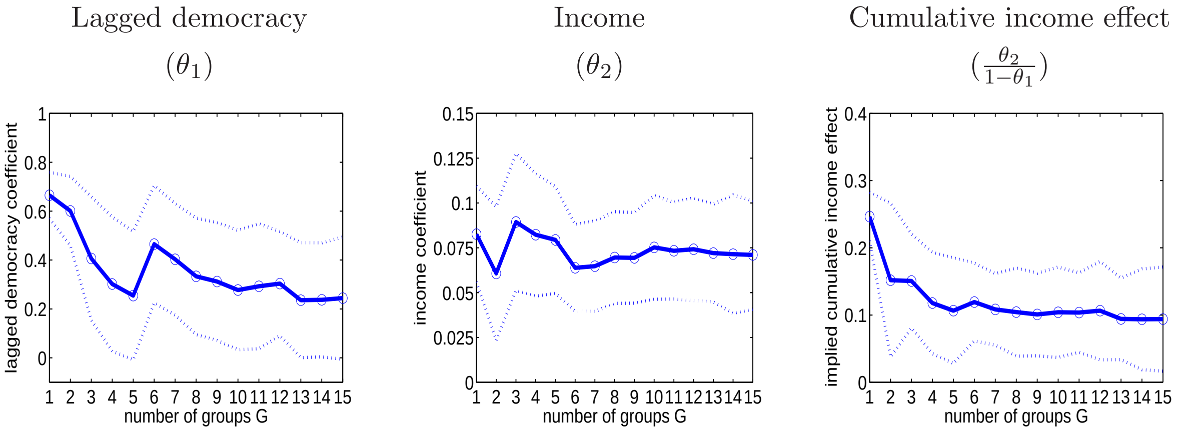
Note that, though statistically significant, the cumulative income effect is quantitatively small. Moreover, we show in the supplementary appendix that the association between income and democracy disappears in a specification that combines both time-varying grouped effects and time-invariant country-specific effects, as in model (5). Hence, in this specification which nests the one in Acemoglu *et al.* (2008), the income effect is not statistically different from zero.

Grouped patterns. The GFE estimates of the unobserved determinants of democracy reveal heterogeneous, time-varying patterns. The upper panel of Figure 2 shows the estimates of group membership by country on a World map, when $G = 4$. The bottom panel shows the parameter estimates

[16] All data in this section are taken from the files of Acemoglu *et al.* (2008): http://economics.mit.edu/files/5000

[17] All estimates are computed using Algorithm 2. We performed extensive checks of numerical accuracy, some of which are described in the supplementary appendix. Stata codes to replicate the results are available as supplementary material.

Figure 1: Coefficients of income and lagged democracy



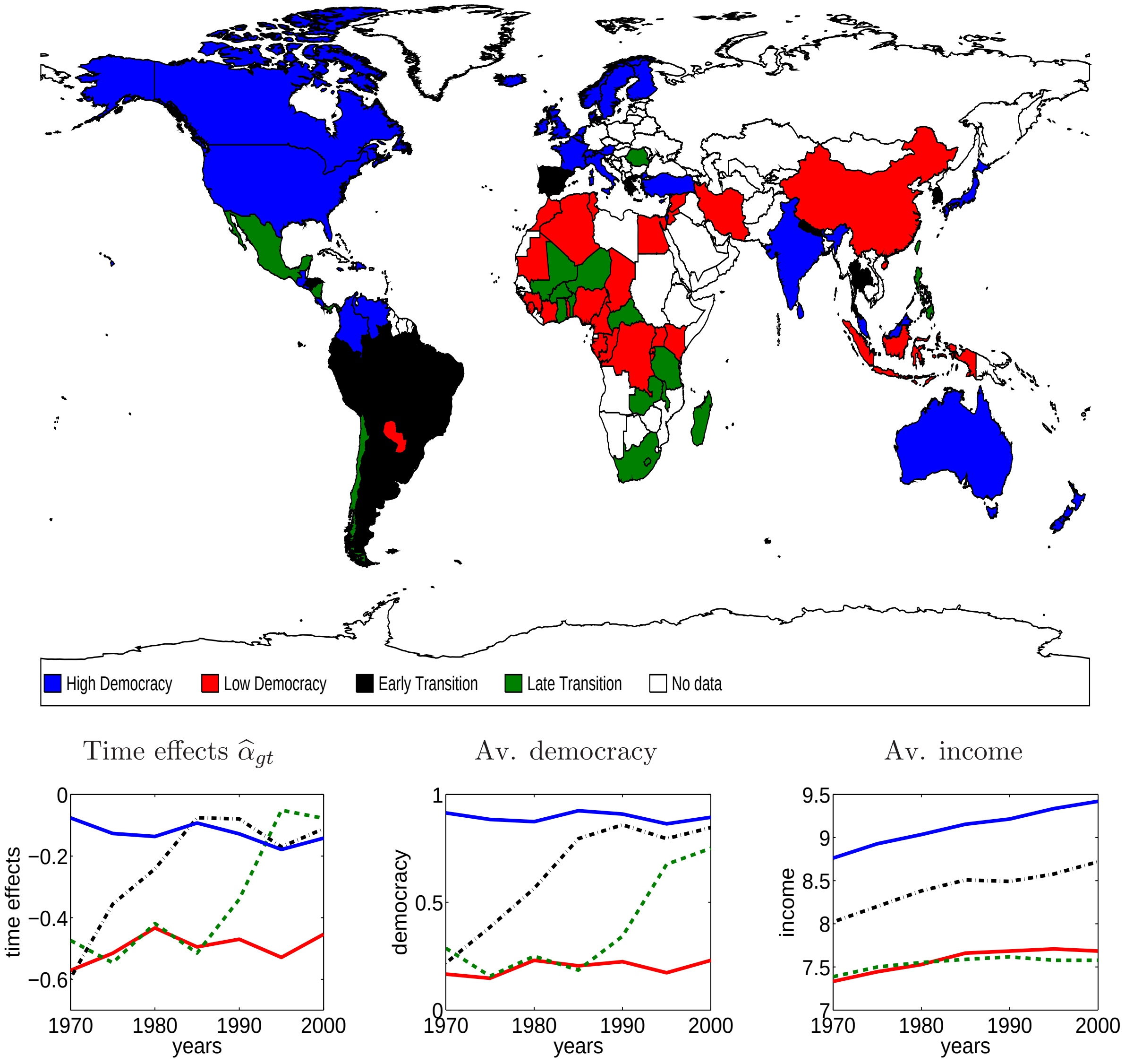
Note: Balanced panel from Acemoglu et al. (2008). The x-axis shows the number of groups G used in estimation, the y-axis reports parameter values. 95%*-confidence intervals clustered at the country level are shown in dashed lines. Confidence intervals are based on bootstrapped standard errors (*100 *replications). Details on the computation are provided in the supplementary appendix.*

$\widehat{\alpha}_{gt}$, and average democracy and lagged log-GDP per capita by group over time. In the supplementary appendix we report the results of specifications with different choices of number of groups and variables used, as well as estimates that also take into account time-invariant country-specific fixed-effects or heterogeneity in coefficients. The qualitative classification shown on the map is remarkably robust across these different specifications.

Figure 2 shows that two of the four groups experience stable paths of democracy over time, albeit at very different levels, while the other two show upward-sloping profiles. Group 1, which we refer to as the "high-democracy" group, mostly contains high-income, high-democracy countries. It includes the US and Canada, most of Continental Europe, Japan and Australia, but also India and Costa Rica. Group 2, which we refer to as "low-democracy", mostly includes low-income, low-democracy countries: a large share of North and Central Africa, China, and Iran, among others. Groups 1 and 2, which together account for 59 of the 90 countries, are broadly consistent with an additive fixed-effects representation, as their grouped effects $\widehat{\alpha}_{1t}$ and $\widehat{\alpha}_{2t}$ are approximately parallel over time. In addition, the graph of average income by group shows that group membership is strongly correlated with log-GDP per capita, consistently with the presence of an upward omitted variable bias in the cross-sectional regression of democracy on income.

While the first two groups of countries are consistent with a fixed-effects model, the other two are not. Group 3 ("early transition") experiences a marked increase in democracy in the first part of the sample period: its mean Freedom House score increases from .20 in 1970 to almost .90 in 1990. This group includes a large share of Latin America, Greece, Spain and Portugal, Thailand and South Korea,

Figure 2: Patterns of heterogeneity, $G = 4$



Note: See the notes to Figure 1. On the bottom panel, the left graph shows the group-specific time effects $\widehat{\alpha}_{gt}$. The other two graphs show the group-specific averages of democracy and lagged log-GDP per capita, respectively. Calendar years (1970-2000) are shown on the x-axis. Light solid lines correspond to Group 1 ("high-democracy"), dark solid lines to Group 2 ("low-democracy"), light dashed lines to Group 3 ("early transition"), and dark dashed lines to Group 4 ("late transition"). The top panel shows group membership. The list of countries by group is given in the supplementary appendix.

in total 13 countries, with an intermediate level of GDP per capita. Group 4 ("late transition") makes a later transition to democracy: its average Freedom House score increases from .20 to .75 between 1985 and 2000. This group includes 18 countries, among which are a large part of West and South Africa, Chile, Romania, and Philippines. These are low-income countries, whose GDP per capita is similar on average to that of the "low-democracy" group (Group 2).

Note that the time patterns and group membership in Figure 2 are estimated from the panel data, and not driven by modelling assumptions other than the grouped structure. In particular, nothing in our framework imposes that time patterns are smooth over time. Moreover, group membership is not assumed to have a particular spatial structure, so the geographic correlation apparent on the map is purely a result of estimation.

Discussion. Overall, the evidence obtained suggests that the effect of income on democracy is perhaps zero, or in any case quantitatively small, in line with the conclusions of Acemoglu *et al.* (2008). At the same time, our analysis highlights the presence of clustering in the evolution of political outcomes: while a substantial share of the world seems to have experienced stable parallel political patterns during the period, roughly one third of the sample has experienced steep upward transitions.

These results raise interesting questions: which factors explain democratic transitions, and importantly, why we observe groups of countries making transitions at similar points in time. In the supplementary appendix we present a first attempt at explaining why these groups of countries have evolved so differently, by regressing group membership on various determinants (including historical measures). We see our framework as providing a starting point to assess how well different theories of democratization fit the political and economic evolution of countries over time.

5 Conclusion

Grouped fixed-effects (GFE) offers a flexible yet parsimonious approach to model unobserved heterogeneity. The approach delivers estimates of common regression parameters, together with interpretable estimates of group-specific time patterns and group membership. The framework allows for strictly exogenous covariates and lagged outcomes. It also easily accommodates unit-specific fixed-effects in addition to the time-varying grouped patterns, and grouped heterogeneity in coefficients. Importantly, the relationship between group membership and observed covariates is left unrestricted.

The GFE approach should be useful in applications where time-varying grouped effects may be present in the data. As a first example, the empirical analysis of the evolution of democracy shows evidence of a clustering of political regimes and transitions. More generally, GFE should be well-suited in difference-in-difference designs, as a way to relax parallel trend assumptions. Other potential applications include models of social interactions and spatial dependence where the reference groups or the spatial weights matrix are estimated from the panel data.

The extension to nonlinear models is a natural next step. While it is possible to define GFE estimators in more general models (see for example equation (9)), the analysis raises statistical challenges. One area of applications is static or dynamic discrete choice modelling, where a discrete specification of unobserved heterogeneity may be appealing (Kasahara and Shimotsu, 2009, Browning and Carro, 2013). See Saggio (2012) for a first attempt in this direction.

Lastly, another interesting extension is to relax the assumption that there is a finite number of well-separated groups in the population. As an alternative approach, one could view the grouped model as an approximation to the underlying data generating process, and characterize the statistical properties of GFE as the number of groups G increases with the two dimensions of the panel.

References

[1] Acemoglu, D., S. Johnson, J. Robinson, and P. Yared (2008): “Income and Democracy,” *American Economic Review*, 98, 808–842.

[2] Ahlquist, J., and E. Wibbels (2012): “Riding the Wave: World Trade and Factor-Based Models of Democratization,” to appear in *American Journal of Political Science.*

[3] Andrews, D. (1984): “Non-Strong Mixing Autoregressive Processes,” *Journal of Applied Probability*, 21, 930–934.

[4] Arellano, M., and J. Hahn (2007): “Understanding Bias in Nonlinear Panel Models: Some Recent Developments,”. In: R. Blundell, W. Newey, and T. Persson (eds.): *Advances in Economics and Econometrics, Ninth World Congress*, Cambridge University Press.

[5] Bai, J. (2003), “Inferential Theory for Factor Models of Large Dimensions,” *Econometrica*, 71, 135–171.

[6] Bai, J. (2009), “Panel Data Models with Interactive Fixed Effects,” *Econometrica*, 77, 1229–1279.

[7] Bai, J., and S. Ng (2002): “Determining the Number of Factors in Approximate Factor Models,” *Econometrica*, 70, 191–221.

[8] Barro, R. J. (1999): “Determinants of Democracy,” *Journal of Political Economy*, 107(6), S158–83.

[9] Bester, A., and C. Hansen (2013): “Grouped Effects Estimators in Fixed Effects Models”, to appear in the *Journal of Econometrics.*

[10] Bickel, P.J., and A. Chen (2009): “A Nonparametric View of Network Models and Newman-Girvan and Other Modularities,” *Proc. Natl. Acad. Sci. USA*, 106, 21068–21073.

[11] Blume, L.E., W.A. Brock, S.N. Durlauf, and Y.M. Ioannides (2011): “Identification of Social Interactions,” in: J. Benhabib, A. Bisin, and M.O. Jackson (Eds.), H*andbook of Social Economics*, Amsterdam: Elsevier Science.

[12] Browning, M., and J. Carro (2013): “Dynamic Binary Outcome Models with Maximal Heterogeneity”, to appear in the *Journal of Econometrics.*

[13] Bryant, P. and Williamson, J. A. (1978): “Asymptotic Behaviour of Classification Maximum Likelihood Estimates,” *Biometrika*, 65, 273–281.

[14] Canova, F. (2004): “Testing for Convergence Clubs in Income per Capita: A Predictive Density Approach,” *International Economic Review*, 45(1), 49–77.

[15] Choi, D.S., P.J. Wolfe, and E.M. Airoldi (2012): "Stochastic Blockmodels with a Growing Number of Classes," *Biometrika*, 99, 273–284.

[16] Forgy, E.W. (1965): "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications," *Biometrics*, 21, 768–769.

[17] Frühwirth-Schnatter, S. (2006): *Finite Mixture and Markov Switching Models*, Springer.

[18] Gleditsch, K.S., and M.D. Ward (2006): "Diffusion and the International Context of Democratization," *International Organization*, 60, 911–933.

[19] Hahn, J., and H. Moon (2010): "Panel Data Models with Finite Number of Multiple Equilibria," *Econometric Theory*, 26(3), 863–881.

[20] Hansen, P., and N. Mladenović (2001): "J-Means: A New Local Search Heuristic for Minimum Sum-of-Squares Clustering," *Pattern Recognition*, 34(2), 405–413.

[21] Hansen, P., N. Mladenović, and J. A. Moreno Pérez (2010): "Variable Neighborhood Search: Algorithms and Applications," *Annals of Operations Research*, 175, 367–407.

[22] Huntington, S.P. (1991): *The Third Wave: Democratization in the Late Twentieth Century*, Norman, OK, and London: University of Oklahoma Press.

[23] Kasahara, H., and K. Shimotsu (2009): "Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices," *Econometrica*, 77(1), 135–175.

[24] Lin, C. C., and S. Ng (2012): "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown", *Journal of Econometric Methods*, 1(1), 42–55.

[25] Lipset, S. M. (1959): "Some Social Requisites of Democracy: Economic Development and Political Legitimacy," *American Political Science Review*, 53(1), 69–105.

[26] McLachlan, G., and D. Peel (2000): *Finite Mixture Models*, Wiley Series in Probabilities and Statistics.

[27] Merlevède, F., Peligrad, M. and E. Rio (2011): "A Bernstein Type Inequality and Moderate Deviations for Weakly Dependent Sequences," *Probability Theory and Related Fields*, 151, 435–474.

[28] Nickel, S. (1981): "Biases in Dynamic Models with Fixed Effects," *Econometrica*, 49, 1417–1426.

[29] Phillips, P.C.B., and D. Sul (2007): "Transition Modelling and Econometric Convergence Tests," *Econometrica*, 75, 1771–1855.

[30] Pollard, D. (1981): “Strong Consistency of K-means Clustering,” *Annals of Statistics*, 9, 135–140.

[31] Pollard, D. (1982): “A Central Limit Theorem for K-Means Clustering,” *Annals of Probability*, 10, 919–926.

[32] Rio, E. (2000): *Théorie Asymptotique des Processus Aléatoires Faiblement Dépendants*, SMAI, Springer.

[33] Saggio, R. (2012): “Discrete Unobserved Heterogeneity in Discrete Choice Panel Data Models,” CEMFI Master Thesis.

[34] Sarafidis, V., and T. Wansbeek (2012): “Cross-sectional Dependence in Panel Data Analysis,” to appear in *Econometric Reviews*.

[35] Steinley, D. (2006): “K-means Clustering: A Half-Century Synthesis,” *Br. J. Math. Stat. Psychol.*, 59, 1–34.

[36] Stock, J., and M. Watson (2002): “Forecasting Using Principal Components from a Large Number of Predictors,” *Journal of the American Statistical Association*, 97, 1167–1179.

[37] Sun, Y. (2005): “Estimation and Inference in Panel Structure Models,” unpublished manuscript.

[38] Tibshirani, R. (1996): “Regression Shrinkage and Selection via the Lasso,” *J. Roy. Statist. Soc. Ser. B*, 58, 267–288.

[39] Townsend, R. M. (1994): “Risk and Insurance in Village India,” *Econometrica*, 62, 539–91.

APPENDIX

A Proof of Theorem 1

Let $\gamma^0 = \{g_1^0, ..., g_N^0\}$ denote the population grouping. Let also $\gamma = \{g_1, ..., g_N\}$ denote any grouping of the cross-sectional units into G groups. Let us define:

$$\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) = \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2. \tag{A1}$$

Note that the GFE estimator minimizes $\widehat{\mathcal{Q}}(\cdot)$ over all $(\theta, \alpha, \gamma) \in \Theta \times \mathcal{A}^{GT} \times \Gamma_G$. Note also that:

$$\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) \quad = \quad \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(v_{it} + x_{it}'\left(\theta^0 - \theta\right) + \alpha^0_{g_i^0 t} - \alpha_{g_i t}\right)^2.$$

We also define the following auxiliary objective function:

$$\widetilde{\mathcal{Q}}(\theta, \alpha, \gamma) \quad = \quad \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(x_{it}'\left(\theta^0 - \theta\right) + \alpha^0_{g_i^0 t} - \alpha_{g_i t}\right)^2 + \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T v_{it}^2.$$

We start by showing the following uniform convergence result.

Lemma A1 *Let Assumption 1.a-1.f hold. Then:*

$$\operatorname*{plim}_{N,T\to\infty} \sup_{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G} \left|\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) - \widetilde{\mathcal{Q}}(\theta, \alpha, \gamma)\right| = 0.$$

Proof.

$$\begin{aligned}
\widehat{\mathcal{Q}}(\theta, \alpha, \gamma) - \widetilde{\mathcal{Q}}(\theta, \alpha, \gamma) \quad &= \quad \frac{2}{NT}\sum_{i=1}^N \sum_{t=1}^T v_{it}\left(x_{it}'\left(\theta^0 - \theta\right) + \alpha^0_{g_i^0 t} - \alpha_{g_i t}\right) \\
&= \quad \left(\frac{2}{NT}\sum_{i=1}^N \sum_{t=1}^T v_{it}x_{it}\right)'\left(\theta^0 - \theta\right) + \frac{2}{NT}\sum_{i=1}^N \sum_{t=1}^T v_{it}\alpha^0_{g_i^0 t} - \frac{2}{NT}\sum_{i=1}^N \sum_{t=1}^T v_{it}\alpha_{g_i t}.
\end{aligned}$$

By Assumption 1.d we have:

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \left\|\frac{1}{T}\sum_{t=1}^T v_{it}x_{it}\right\|^2\right] \leq \frac{M}{T},$$

so it follows from the Cauchy-Schwartz (CS) inequality that $\frac{2}{NT}\sum_{i=1}^N\sum_{t=1}^T v_{it}x_{it} = o_p(1)$. In addition, $\left\|\theta^0 - \theta\right\|$ is bounded by Assumption 1.a.

We next show that $\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T v_{it}\alpha_{g_i t}$ is $o_p(1)$, uniformly on the parameter space. This will imply that $\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T v_{it}\alpha^0_{g_i^0 t} = o_p(1)$. We have:

$$\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T v_{it}\alpha_{g_i t} \quad = \quad \sum_{g=1}^G \left[\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \mathbf{1}\{g_i = g\}v_{it}\alpha_{gt}\right] \quad = \quad \sum_{g=1}^G \left[\frac{1}{T}\sum_{t=1}^T \alpha_{gt}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{g_i = g\}v_{it}\right)\right].$$

Moreover, by the CS inequality and for all $g \in \{1, ..., G\}$:

$$\left(\frac{1}{T}\sum_{t=1}^T \alpha_{gt}\left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{g_i = g\}v_{it}\right)\right)^2 \quad \leq \quad \left(\frac{1}{T}\sum_{t=1}^T \alpha_{gt}^2\right) \times \left(\frac{1}{T}\sum_{t=1}^T \left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{g_i = g\}v_{it}\right)^2\right),$$

where, by Assumption 1.a, $\frac{1}{T}\sum_{t=1}^{T}\alpha_{gt}^2$ is uniformly bounded. Now, note that:

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i=g\}v_{it}\right)^2 &= \frac{1}{TN^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{1}\{g_i=g\}\mathbf{1}\{g_j=g\}\sum_{t=1}^{T}v_{it}v_{jt} \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}v_{it}v_{jt}\right| \\ &\leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(v_{it}v_{jt}\right)\right| + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(v_{it}v_{jt}-\mathbb{E}\left(v_{it}v_{jt}\right)\right)\right|.\end{aligned}$$

By Assumption 1.e: $\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(v_{it}v_{jt}\right)\right| \leq \frac{M}{N}$. Moreover, by the CS inequality:

$$\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|\frac{1}{T}\sum_{t=1}^{T}\left(v_{it}v_{jt}-\mathbb{E}\left(v_{it}v_{jt}\right)\right)\right|\right)^2 \leq \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{1}{T}\sum_{t=1}^{T}\left(v_{it}v_{jt}-\mathbb{E}\left(v_{it}v_{jt}\right)\right)\right)^2,$$

which is bounded in expectation by M/T by Assumption 1.f.

This shows that $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}v_{it}\alpha_{g_it}$ is uniformly $o_p(1)$, and ends the proof of Lemma A1.

■

The following result shows that $\widetilde{\mathcal{Q}}(\cdot)$ is uniquely minimized at true values.

Lemma A2 *For all $(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G$:*

$$\widetilde{\mathcal{Q}}(\theta,\alpha,\gamma)-\widetilde{\mathcal{Q}}\left(\theta^0,\alpha^0,\gamma^0\right)\geq\widehat{\rho}\left\|\theta-\theta^0\right\|^2,$$

where $\widehat{\rho}$ is given by Assumption 1.g.

Proof. Let us denote, for every grouping $\gamma=\{g_1,...,g_N\}$:

$$\Sigma(\gamma)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\overline{x}_{g_i^0\wedge g_i,t}\right)\left(x_{it}-\overline{x}_{g_i^0\wedge g_i,t}\right)'.$$

We have, using standard least-squares algebra:

$$\begin{aligned}\widetilde{\mathcal{Q}}(\theta,\alpha,\gamma)-\widetilde{\mathcal{Q}}\left(\theta^0,\alpha^0,\gamma^0\right) &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^0-\theta\right)+\alpha^0_{g_i^0t}-\alpha_{g_it}\right)^2 \geq \left(\theta^0-\theta\right)'\Sigma(\gamma)\left(\theta^0-\theta\right) \\ &\geq \min_{\gamma\in\Gamma_G}\left(\theta^0-\theta\right)'\Sigma(\gamma)\left(\theta^0-\theta\right) \geq \left(\min_{\gamma\in\Gamma_G}\widehat{\rho}(\gamma)\right)\left\|\theta^0-\theta\right\|^2,\end{aligned}$$

where $\min_{\gamma\in\Gamma_G}\widehat{\rho}(\gamma)$ is asymptotically bounded away from zero by Assumption 1.g.

■

To show that $\widehat{\theta}$ is consistent for θ^0, note that, by Lemma A1 and by the definition of the GFE estimator:

$$\widetilde{\mathcal{Q}}\left(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}\right) = \widehat{\mathcal{Q}}\left(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}\right)+o_p(1) \leq \widehat{\mathcal{Q}}\left(\theta^0,\alpha^0,\gamma^0\right)+o_p(1) = \widetilde{\mathcal{Q}}\left(\theta^0,\alpha^0,\gamma^0\right)+o_p(1). \tag{A2}$$

So, by Lemma A2 and Assumption 1.g: $\left\|\widehat{\theta}-\theta^0\right\|^2=o_p(1)$.

Lastly, to show convergence in quadratic mean of the estimated unit-specific effects, note that:

$$\begin{aligned}\left|\widetilde{\mathcal{Q}}\left(\widehat{\theta},\widehat{\alpha},\widehat{\gamma}\right)-\widetilde{\mathcal{Q}}\left(\theta^0,\widehat{\alpha},\widehat{\gamma}\right)\right| &= \left|\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}'\left(\theta^0-\widehat{\theta}\right)\left[x_{it}'\left(\theta^0-\widehat{\theta}\right)+2\left(\alpha^0_{g_i^0t}-\widehat{\alpha}_{\widehat{g}_it}\right)\right]\right| \\ &\leq \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|x_{it}\right\|^2\times\left\|\theta^0-\widehat{\theta}\right\|^2+\left(4\sup_{\alpha_t\in\mathcal{A}}|\alpha_t|\right)\times\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\|x_{it}\right\|\times\left\|\theta^0-\widehat{\theta}\right\|,\end{aligned}$$

which is $o_p(1)$ by Assumptions 1.a and 1.b, and by consistency of $\widehat{\theta}$. Combining with (A2) we obtain: $\widetilde{\mathcal{Q}}\left(\theta^0, \widehat{\alpha}, \widehat{\gamma}\right) \leq \widetilde{\mathcal{Q}}\left(\theta^0, \alpha^0, \gamma^0\right) + o_p(1)$, from which it follows that: $\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2 = o_p(1)$.

This completes the proof of Theorem 1.

B Proof of Theorem 2

We first establish that $\widehat{\alpha}$ is consistent for α^0. Because the objective function is invariant to re-labelling of the groups, we show consistency with respect to the Hausdorff distance d_H in $\mathbb{R}^{GT}$, defined by:

$$d_H\left(a,b\right)^2 = \max\left\{ \max_{g\in\{1,...,G\}} \left(\min_{\widetilde{g}\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^T \left(a_{\widetilde{g}t} - b_{gt}\right)^2 \right), \max_{\widetilde{g}\in\{1,...,G\}} \left(\min_{g\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^T \left(a_{\widetilde{g}t} - b_{gt}\right)^2 \right) \right\}.$$

We have the following result.[18]

Lemma B3 *Let Assumptions 1.a-1.g, and 2.a-2.b hold. Then, as N and T tend to infinity:*

$$d_H\left(\widehat{\alpha}, \alpha^0\right) \xrightarrow{p} 0.$$

Proof.

We study the two terms in the $\max\{\cdot,\cdot\}$ in turn.

• We first show that, for all $g \in \{1,...,G\}$:

$$\min_{\widetilde{g}\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^T \left(\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt}\right)^2 \xrightarrow{p} 0. \tag{B3}$$

Let $g \in \{1,...,G\}$. We have:

$$\frac{1}{NT}\sum_{i=1}^N \left(\min_{\widetilde{g}\in\{1,...,G\}} \sum_{t=1}^T \mathbf{1}\{g_i^0 = g\} \left(\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt}\right)^2 \right) = \left(\frac{1}{N}\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\} \right) \left(\min_{\widetilde{g}\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^T \left(\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt}\right)^2 \right).$$

By Assumption 2.a it is thus enough to show that, for all g, as N and T tend to infinity:

$$\frac{1}{NT}\sum_{i=1}^N \left(\min_{\widetilde{g}\in\{1,...,G\}} \sum_{t=1}^T \mathbf{1}\{g_i^0 = g\} \left(\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt}\right)^2 \right) \xrightarrow{p} 0.$$

Now:

$$\begin{aligned} \frac{1}{NT}\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\} \left(\min_{\widetilde{g}\in\{1,...,G\}} \sum_{t=1}^T \left(\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt}\right)^2 \right) &\leq \frac{1}{NT}\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\} \left(\sum_{t=1}^T \left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{gt}\right)^2 \right) \\ &\leq \frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \left(\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}\right)^2, \end{aligned}$$

which is $o_p(1)$ by Theorem 1. Hence (B3) follows.

• Let us define, for all $g \in \{1,...,G\}$:

$$\sigma(g) = \underset{\widetilde{g}\in\{1,...,G\}}{\text{argmin}} \frac{1}{T}\sum_{t=1}^T \left(\widehat{\alpha}_{\widetilde{g}t} - \alpha^0_{gt}\right)^2.$$

[18] Note that group separation (Assumption 2.b) is assumed to show Lemma B3. Proving consistency of the group-specific time effects absent this assumption would require different arguments.

We start by showing that $\sigma : \{1,...,G\} \to \{1,...,G\}$ is one-to-one, with probability approaching one as T tends to infinity. Let $g \neq \widetilde{g}$. By the triangle inequality we have:

$$\left(\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(g)t}-\widehat{\alpha}_{\sigma(\widetilde{g})t}\right)^2\right)^{\frac{1}{2}} \geq \left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\right)^{\frac{1}{2}} - \left(\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(g)t}-\alpha_{gt}^0\right)^2\right)^{\frac{1}{2}} - \left(\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(\widetilde{g})t}-\alpha_{\widetilde{g}t}^0\right)^2\right)^{\frac{1}{2}},$$

where the right-hand-side of this inequality converges in probability to $(c_{g,\widetilde{g}})^{\frac{1}{2}}$ by Assumption 2.b and equation (B3). It thus follows that, with probability approaching one, $\sigma(g) \neq \sigma(\widetilde{g})$ for all $g \neq \widetilde{g}$. Thus σ admits a well-defined inverse σ^{-1}.

Now, with probability approaching one we have, for all $\widetilde{g} \in \{1,...,G\}$:

$$\min_{g\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widetilde{g}t}-\alpha_{gt}^0\right)^2 \leq \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widetilde{g}t}-\alpha_{\sigma^{-1}(\widetilde{g})t}^0\right)^2 = \min_{h\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{ht}-\alpha_{\sigma^{-1}(\widetilde{g})t}^0\right)^2 \xrightarrow{p} 0,$$

where we have used (B3), and the fact that $\widetilde{g} = \sigma\left[\sigma^{-1}(\widetilde{g})\right]$. Combining with (B3) completes the proof.

■

The proof of Lemma B3 shows that there exists a permutation $\sigma : \{1,...,G\} \to \{1,...,G\}$ such that:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\sigma(g)t}-\alpha_{gt}^0\right)^2 \xrightarrow{p} 0.$$

By simple relabelling of the elements of $\widehat{\alpha}$ we may take $\sigma(g) = g$. We adopt this convention in the rest of the proof. For any $\eta > 0$, we let $\mathcal{N}_\eta$ denote the set of parameters $(\theta, \alpha) \in \Theta \times \mathcal{A}^{GT}$ that satisfy $\left\|\theta - \theta^0\right\|^2 < \eta$ and $\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha_{gt}^0\right)^2 < \eta$ for all $g \in \{1,...,G\}$. We have the following result.

Lemma B4 *For $\eta > 0$ small enough we have, for all $\delta > 0$ and as N and T tend to infinity:*

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} = o_p\left(T^{-\delta}\right).$$

Proof.

Note that, from the definition of $\widehat{g}_i(\cdot)$ we have, for all $g \in \{1,...,G\}$:

$$\mathbf{1}\{\widehat{g}_i(\theta,\alpha) = g\} \leq \mathbf{1}\left\{\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{gt}\right)^2 \leq \sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{g_i^0t}\right)^2\right\},$$

so:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(\theta,\alpha) \neq g_i^0\} &= \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\{\widehat{g}_i(\theta,\alpha) = g\} \\ &\leq \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\left\{\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{gt}\right)^2 \leq \sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{g_i^0t}\right)^2\right\}}_{=Z_{ig}(\theta,\alpha)}.\end{aligned}$$

We start by bounding $Z_{ig}(\theta,\alpha)$, for all $(\theta,\alpha) \in \mathcal{N}_\eta$, by a quantity that does not depend on (θ,α). To proceed note that, for all (θ,α) and all i:

$$\begin{aligned}Z_{ig}(\theta,\alpha) &= \mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\left\{\sum_{t=1}^{T}\left(\alpha_{g_i^0t}-\alpha_{gt}\right)\left(v_{it}+x_{it}'\left(\theta^0-\theta\right)+\alpha_{g_i^0t}^0-\frac{\alpha_{gt}+\alpha_{g_i^0t}}{2}\right) \leq 0\right\} \\ &\leq \max_{\widetilde{g}\neq g}\mathbf{1}\left\{\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)\left(v_{it}+x_{it}'\left(\theta^0-\theta\right)+\alpha_{\widetilde{g}t}^0-\frac{\alpha_{gt}+\alpha_{\widetilde{g}t}}{2}\right) \leq 0\right\}.\end{aligned}$$

Let us now define:

$$A_T = \left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)\left(v_{it}+x_{it}'\left(\theta^0-\theta\right)+\alpha_{\widetilde{g}t}^0-\frac{\alpha_{gt}+\alpha_{\widetilde{g}t}}{2}\right)-\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)\left(v_{it}+\alpha_{\widetilde{g}t}^0-\frac{\alpha_{gt}^0+\alpha_{\widetilde{g}t}^0}{2}\right)\right|.$$

As we have:

$$\begin{aligned} A_T \quad &\leq \quad \left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)v_{it}-\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)v_{it}\right|+\left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)x_{it}'\left(\theta^0-\theta\right)\right| \\ &+\left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)\left(\alpha_{\widetilde{g}t}^0-\frac{\alpha_{gt}+\alpha_{\widetilde{g}t}}{2}\right)-\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)\left(\alpha_{\widetilde{g}t}^0-\frac{\alpha_{gt}^0+\alpha_{\widetilde{g}t}^0}{2}\right)\right|, \end{aligned}$$

it is easy to show using the CS inequality that, for $(\theta,\alpha)\in\mathcal{N}_\eta$:

$$A_T \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}}+TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)+TC_3\sqrt{\eta},$$

where C_1, C_2 and C_3 are constants, independent of η and T.

We thus obtain that:

$$\begin{aligned} Z_{ig}(\theta,\alpha) \quad &\leq \quad \max_{\widetilde{g}\neq g}\mathbf{1}\left\{\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)\left(v_{it}+\alpha_{\widetilde{g}t}^0-\frac{\alpha_{gt}^0+\alpha_{\widetilde{g}t}^0}{2}\right)\right. \\ &\left.\leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}}+TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)+TC_3\sqrt{\eta}\right\}. \end{aligned}$$

Noting that the right-hand side of this inequality does not depend on (θ,α), it follows that: $\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} Z_{ig}(\theta,\alpha)\leq \widetilde{Z}_{ig}$, where:

$$\begin{aligned} \widetilde{Z}_{ig} \quad &= \quad \max_{\widetilde{g}\neq g}\mathbf{1}\left\{\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)v_{it}\leq -\frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)^2+TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}}\right. \\ &\left.+TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)+TC_3\sqrt{\eta}\right\}. \qquad \text{(B4)} \end{aligned}$$

As a result:

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i\left(\theta,\alpha\right)\neq g_i^0\}\leq\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Z}_{ig}. \qquad \text{(B5)}$$

Fix $\widetilde{M}>\max(\sqrt{M},M^*)$, where M and M^* are given by Assumptions 1 and 2.e, respectively. Note that $\mathbb{E}(v_{it}^2)\leq\sqrt{M}$. We have, using standard probability algebra and for all g:

$$\begin{aligned} \Pr\left(\widetilde{Z}_{ig}=1\right) \quad &\leq \quad \sum_{\widetilde{g}\neq g}\Pr\left(\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)v_{it}\leq -\frac{1}{2}\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)^2+TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}}\right. \\ &\left.+TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)+TC_3\sqrt{\eta}\right) \\ &\leq \quad \sum_{\widetilde{g}\neq g}\left[\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\geq\widetilde{M}\right)+\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)^2\leq\frac{c_{g,\widetilde{g}}}{2}\right)+\Pr\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\geq\widetilde{M}\right)\right. \\ &\left.+\Pr\left(\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)v_{it}\leq -T\frac{c_{g,\widetilde{g}}}{4}+TC_1\sqrt{\eta}\sqrt{\widetilde{M}}+TC_2\sqrt{\eta}\widetilde{M}+TC_3\sqrt{\eta}\right)\right]. \end{aligned}$$

(B6)

To end the proof of Lemma B4, we rely on the use of exponential inequalities for dependent processes. Specifically, we use the following result, which is a direct consequence of Theorem 6.2 in Rio (2000).

Lemma B5 *Let z_t be a strongly mixing process with zero mean, with strong mixing coefficients $\alpha[t] \leq e^{-at^{d_1}}$, and with tail probabilities $\Pr(|z_t| > z) \leq e^{1-\left(\frac{z}{b}\right)^{d_2}}$, where a, b, d_1, and d_2 are positive constants. Then, for all $z > 0$ we have, for all $\delta > 0$:*

$$T^{\delta} \Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \overset{T\to\infty}{\to} 0.$$

Proof. Let $s^2 = \sup_{t\geq 1}\left(\sum_{s\geq 1}|\mathbb{E}(z_t z_s)|\right)$. Note that $s^2 < \infty$ under the condition of Lemma B5.[19] Let also $d = \frac{d_1 d_2}{d_1+d_2}$. By evaluating inequality (1.7) in Merlevède *et al.* (2011) at $\lambda = T\frac{z}{4}$ and $r = T^{\frac{1}{2}}$, we obtain that there exists a constant $f > 0$ independent of T such that, for all $z > 0$ and $T \geq 1$:

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T} z_t\right| \geq z\right) \leq 4\left(1 + T^{\frac{1}{2}}\frac{z^2}{16s^2}\right)^{-\frac{1}{2}T^{\frac{1}{2}}} + \frac{16f}{z}\exp\left(-a\left(T^{\frac{1}{2}}\frac{z}{4b}\right)^{d}\right).$$

Lemma B5 directly follows.

■

We now bound the last three terms on the right-hand side of (B6).

• By Assumptions 1.a and 2.b we have $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2\right] = c_{g,\widetilde{g}}$. So for T large enough we have:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2\right] \geq \frac{2c_{g,\widetilde{g}}}{3}.$$

Applying Lemma B5 to $z_t = \left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2 - \mathbb{E}\left[\left(\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\right)^2\right]$, which satisfies appropriate mixing and tail conditions by Assumptions 1.a and 2.c, and taking $z = \frac{c_{g,\widetilde{g}}}{6}$ yields, for all $\delta > 0$ and as T tends to infinity:

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0\right)^2 \leq \frac{c_{g,\widetilde{g}}}{2}\right) = o\left(T^{-\delta}\right).$$

• Similarly, for the third term on the right-hand side of (B6), applying Lemma B5 to $z_t = v_{it}^2 - \mathbb{E}(v_{it}^2)$ and taking $z = \widetilde{M} - \sqrt{M}$ yields:

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T} v_{it}^2 \geq \widetilde{M}\right) = o\left(T^{-\delta}\right)$$

for all $\delta > 0$. Note that $\{v_{it}^2\}_t$ is strongly mixing as $\{v_{it}\}_t$ is strongly mixing by Assumption 2.c.

• Lastly, to bound the fourth term on the right-hand side of (B6) we denote as c the minimum of $c_{g,\widetilde{g}}$ over all $g \neq \widetilde{g}$ and we take:

$$\eta \leq \left(\frac{c}{8\left(C_1\sqrt{\widetilde{M}} + C_2\widetilde{M} + C_3\right)}\right)^2. \tag{B7}$$

Note that this upper bound on η does not depend on T.

[19] This is a consequence of the fact that $|\operatorname{Cov}(X,Y)|$ can be bounded in terms of the strong mixing coefficient $\alpha[X,Y]$ and the quantile functions Q_X and Q_Y (see, e.g., Theorem 1.1 in Rio, 2000).

Taking η satisfying (B7) yields, for all $\widetilde{g} \neq g$:

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it} \leq -\frac{c_{g,\widetilde{g}}}{4}+C_1\sqrt{\eta}\sqrt{\widetilde{M}}+C_2\sqrt{\eta}\widetilde{M}+C_3\sqrt{\eta}\right) \leq \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it} \leq -\frac{c_{g,\widetilde{g}}}{8}\right).$$

Now, by Assumption 2.c the process $\left\{\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}\right\}_t$ has zero mean, and is strongly mixing with faster-than-polynomial decay rate. Moreover, for all i, t, and $m>0$:

$$\Pr\left(\left|\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}\right|>m\right) \leq \Pr\left(|v_{it}|>\frac{m}{2\sup\limits_{\alpha_t\in\mathcal{A}}|\alpha_t|}\right),$$

so $\left\{\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}\right\}_t$ also satisfies the tail condition of Assumption 2.d, albeit with a different constant $\widetilde{b}>0$ instead of $b>0$.

Lastly, applying Lemma B5 with $z_t=\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}$ and taking $z=\frac{c_{g,\widetilde{g}}}{8}$ yields:

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it} \leq -\frac{c_{g,\widetilde{g}}}{8}\right)=o\left(T^{-\delta}\right). \tag{B8}$$

Note that the above upper bounds on the probabilities do not depend on i and g. Combining results we thus obtain, using (B6) and Assumption 2.e, that for η satisfying (B7) and for all $\delta>0$:

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\Pr\left(\widetilde{Z}_{ig}=1\right) \leq G(G-1)\sup_{i\in\{1,..,N\}}\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\geq\widetilde{M}\right)+o\left(T^{-\delta}\right) = o\left(T^{-\delta}\right). \tag{B9}$$

Sufficient conditions for Assumption 2.e are given at the end of Appendix B.

To complete the proof of Lemma B4 note that, for η that satisfies (B7) we have, for all $\delta>0$ and all $\varepsilon>0$:

$$\Pr\left(\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)\neq g^0_i\}>\varepsilon T^{-\delta}\right) \leq \Pr\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Z}_{ig}>\varepsilon T^{-\delta}\right) \leq \frac{\mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Z}_{ig}\right)}{\varepsilon T^{-\delta}} = o(1),$$

where we have used (B5), the Markov inequality, and (B9), respectively.

This ends the proof of Lemma B4.

■

We now prove the three parts of Theorem 2, and derive asymptotic results for $\widehat{\theta}$, $\widehat{\alpha}$, and $\widehat{g}_i$ in turn.

Properties of $\widehat{\theta}$. Let us denote:[20]

$$\widehat{Q}(\theta,\alpha)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x'_{it}\theta-\alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2, \tag{B10}$$

and:

$$\widetilde{Q}(\theta,\alpha)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x'_{it}\theta-\alpha_{g^0_it}\right)^2. \tag{B11}$$

Note that $\widehat{Q}(\cdot)$ is minimized at $\left(\widehat{\theta},\widehat{\alpha}\right)$, and that $\widetilde{Q}(\cdot)$ is minimized at $\left(\widetilde{\theta},\widetilde{\alpha}\right)$.

[20] Note that $\widehat{Q}(\theta,\alpha)$ is a concentrated version of $\widehat{\mathcal{Q}}(\theta,\alpha,\gamma)$ that was defined in the proof of Theorem 1.

Let $\eta > 0$ be small enough such that Lemma B4 is satisfied. Using Assumptions 1.a-1.c and Lemma B4 it is easy to see that, for all $\delta > 0$:

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \left|\widehat{Q}(\theta,\alpha) - \widetilde{Q}(\theta,\alpha)\right| = o_p\left(T^{-\delta}\right). \tag{B12}$$

Now, by consistency of $\widehat{\theta}$ (Theorem 1) and $\widehat{\alpha}$ (Lemma B3) we have, as N and T tend to infinity:

$$\Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) \to 0. \tag{B13}$$

Likewise, as $\widetilde{\theta}$ and $\widetilde{\alpha}$ are also consistent under the conditions of Theorem 1 we have:

$$\Pr\left(\left(\widetilde{\theta},\widetilde{\alpha}\right) \notin \mathcal{N}_\eta\right) \to 0. \tag{B14}$$

Combining (B12) and (B13) we have, for all $\delta > 0$ and as N and T tend to infinity:

$$\widehat{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right) = o_p\left(T^{-\delta}\right). \tag{B15}$$

This is because, for every $\varepsilon > 0$:

$$\Pr\left[\left|\widehat{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)\right| > \varepsilon T^{-\delta}\right] \quad\leq\quad \Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right) \notin \mathcal{N}_\eta\right) + \Pr\left[\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\left|\widehat{Q}(\theta,\alpha) - \widetilde{Q}(\theta,\alpha)\right| > \varepsilon T^{-\delta}\right],$$

which is $o(1)$ by (B12) and (B13).

Similarly, combining (B12) and (B14) we obtain:

$$\widehat{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) \quad=\quad o_p\left(T^{-\delta}\right). \tag{B16}$$

Hence, using (B15), (B16), and the definition of $\left(\widetilde{\theta},\widetilde{\alpha}\right)$ and $\left(\widehat{\theta},\widehat{\alpha}\right)$ yields:

$$0 \leq \widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) = \widehat{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widehat{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) + o_p(T^{-\delta}) \leq o_p\left(T^{-\delta}\right).$$

It thus follows that:

$$\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) \quad=\quad o_p\left(T^{-\delta}\right). \tag{B17}$$

Now, using that $\left(\widetilde{\theta},\widetilde{\alpha}\right)$ is a least squares estimator, we obtain:

$$\begin{aligned}\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right) - \widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) \quad&=\quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right) + \widetilde{\alpha}_{g_i^0 t} - \widehat{\alpha}_{g_i^0 t}\right)^2 \qquad\qquad (B18)\\ &\geq\quad \left(\widetilde{\theta}-\widehat{\theta}\right)'\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}-\overline{x}_{g_i^0 t}\right)\left(x_{it}-\overline{x}_{g_i^0 t}\right)'\right)\left(\widetilde{\theta}-\widehat{\theta}\right) \;\geq\; \widehat{\rho}\left\|\widetilde{\theta}-\widehat{\theta}\right\|^2,\end{aligned}$$

where $\widehat{\rho} \xrightarrow{p} \rho > 0$ as a consequence of Assumption 1.g. Hence, $\widetilde{\theta} - \widehat{\theta} = o_p\left(T^{-\delta}\right)$ for all $\delta > 0$. This shows (17).

Properties of $\widehat{\alpha}$. Using (B17) and (B18) above, consistency of $\widehat{\theta}$ and $\widetilde{\theta}$, and Assumption 1.b, we obtain:

$$\frac{1}{NT}\sum_{g=1}^{G}\sum_{i=1}^{N}\mathbf{1}\{g_i^0 = g\}\sum_{t=1}^{T}\left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)^2 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widetilde{\alpha}_{g_i^0 t} - \widehat{\alpha}_{g_i^0 t}\right)^2 = o_p\left(T^{-\delta}\right).$$

Using Assumption 2.a, we thus have for all g:

$$\frac{1}{T}\sum_{t=1}^{T}\left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)^2 \quad=\quad o_p\left(T^{-\delta}\right). \tag{B19}$$

In particular, for all t we have: $\left(\widetilde{\alpha}_{gt} - \widehat{\alpha}_{gt}\right)^2 \leq o_p\left(T^{1-\delta}\right)$. As this holds for all $\delta > 0$ we obtain (18).

Properties of $\widehat{g}_i = \widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)$. Finally we have, by the union bound:

$$\Pr\left(\sup_{i\in\{1,\ldots,N\}}\left|\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)-g_i^0\right|>0\right) \quad\leq\quad \Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right)\notin\mathcal{N}_\eta\right)+N\sup_{i\in\{1,\ldots,N\}}\Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right)\in\mathcal{N}_\eta\,,\,\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)\neq g_i^0\right).$$

Now we have, taking η such that (B7) is satisfied: $\Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right)\notin\mathcal{N}_\eta\right)=o(1)$. Moreover, by the proof of Lemma B4 we have: $\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\mathbf{1}\{\widehat{g}_i\left(\theta,\alpha\right)\neq g_i^0\}\leq\sum_{g=1}^G\widetilde{Z}_{ig}$, where $\widetilde{Z}_{ig}$ is given by (B4). Hence:

$$\begin{aligned}
N\sup_{i\in\{1,\ldots,N\}}\Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right)\in\mathcal{N}_\eta,\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)\neq g_i^0\right) \quad &= \quad N\sup_{i\in\{1,\ldots,N\}}\mathbb{E}\left[\mathbf{1}\left\{\left(\widehat{\theta},\widehat{\alpha}\right)\in\mathcal{N}_\eta\right\}\mathbf{1}\left\{\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)\neq g_i^0\right\}\right]\\
&\leq \quad N\sup_{i\in\{1,\ldots,N\}}\mathbb{E}\left[\mathbf{1}\left\{\left(\widehat{\theta},\widehat{\alpha}\right)\in\mathcal{N}_\eta\right\}\sum_{g=1}^G\widetilde{Z}_{ig}\right]\\
&\leq \quad N\sup_{i\in\{1,\ldots,N\}}\sum_{g=1}^G\Pr\left(\widetilde{Z}_{ig}=1\right) \;=\; o\left(NT^{-\delta}\right) \qquad \text{for all } \delta>0,
\end{aligned}$$

where the last equality is obtained similarly as (B9) under Assumption 2.e.

This implies (16), and completes the proof of Theorem 2.

Sufficient conditions for Assumption 2.e. We have the following result.

Proposition B1 *Suppose that either of the following two conditions holds:*

1. *Assumption 1.b holds, and $\{\|x_{it}\|\}_t$ satisfies the mixing and tail conditions of Assumptions 2.c and 2.d.*
2. *Assumptions 1.a, 1.c, and 2.c-2.d, hold. Moreover, $x_{it}=(y_{i,t-1},\widetilde{x}_{it}')'$, and $\theta=\left(\rho,\theta_1'\right)'$, where $|\rho^0|<1$, $\widetilde{x}_{it}$ satisfy Assumption 2.e, and, for all constants $F>0$:*

$$\sup_{i\in\{1,..,N\}}\Pr\left(|y_{i0}|\geq FT\right)=o\left(T^{-\delta}\right)\ \textit{for all}\ \delta>0. \tag{B20}$$

Note that (B20) requires that the distribution of y_{i0} has thinner-than-polynomial tails.

Then there exists a constant $M^*>0$ such that, as N,T tend to infinity:

$$\sup_{i\in\{1,..,N\}}\Pr\left(\frac{1}{T}\sum_{t=1}^T\|x_{it}\|\geq M^*\right)=o\left(T^{-\delta}\right)\ \textit{for all}\ \delta>0.$$

Proof.

Let us first suppose Part 1. Note that $\mathbb{E}(\|x_{it}\|)\leq\sqrt{M}$ by Assumption 1.b. Let $M^*>\sqrt{M}$. The result comes from applying Lemma B5 to $z_t=\|x_{it}\|-\mathbb{E}(\|x_{it}\|)$ and taking $z=M^*-\sqrt{M}$.

Let us then suppose Part 2. By assumption, $\sup_{i\in\{1,..,N\}}\Pr\left(\frac{1}{T}\sum_{t=1}^T\|\widetilde{x}_{it}\|\geq M^*\right)=o\left(T^{-\delta}\right)$. Moreover:

$$y_{it}=\sum_{s=0}^{t-1}\left(\rho^0\right)^s\left(\widetilde{x}_{i,t-s}'\theta_1^0+\alpha_{g_i^0,t-s}^0+v_{i,t-s}\right)+\left(\rho^0\right)^t y_{i0}.$$

So, for all i:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T|y_{i,t-1}| \quad &\leq \quad \frac{1}{T}\sum_{t=1}^T\left[\sum_{s=0}^{t-2}\left(\left|\rho^0\right|^s\left(\|\widetilde{x}_{i,t-1-s}\|\,\|\theta_1^0\|+\left|\alpha_{g_i^0,t-1-s}^0\right|+|v_{i,t-1-s}|\right)\right)+\left|\rho^0\right|^{t-1}|y_{i0}|\right]\\
&\leq \quad \left(\frac{1}{1-|\rho^0|}\right)\left[\frac{1}{T}\sum_{m=1}^{T-1}\left(\|\widetilde{x}_{im}\|\,\|\theta_1^0\|+\left|\alpha_{g_i^0m}^0\right|+|v_{im}|\right)+\frac{1}{T}|y_{i0}|\right],
\end{aligned}$$

where we have used the change-in-variables $m = t - 1 - s$. Taking M^* large enough, Assumptions 1.a, 1.c, 2.c, 2.d, and Lemma B5 imply that:

$$\sup_{i \in \{1,..,N\}} \Pr\left(\frac{1}{T} \sum_{m=1}^{T-1} \left(\|\widetilde{x}_{im}\| \left\|\theta_1^0\right\| + \left|\alpha^0_{g_i^0 m}\right| + |v_{im}| \right) \geq \frac{M^*(1 - \left|\rho^0\right|)}{2} \right) = o(T^{-\delta}).$$

Moreover, for this M^*, (B20) implies that:

$$\sup_{i \in \{1,..,N\}} \Pr\left(\frac{1}{T} |y_{i0}| \geq \frac{M^*(1 - \left|\rho^0\right|)}{2} \right) = o(T^{-\delta}).$$

This concludes the proof.

■