Grouped Patterns of Heterogeneity in Panel Data[*]

Stéphane Bonhomme
CEMFI, Madrid

Elena Manresa
CEMFI, Madrid

December 2012

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. We rely on recent advances in the clustering literature for fast and efficient computation. We provide conditions under which our estimator is higher-order unbiased as both dimensions of the panel tend to infinity, and develop inference methods. We apply our approach to study the link between income and democracy across countries, while allowing for grouped patterns of unobserved heterogeneity. The results shed new light on the evolution of political and economic outcomes of countries.

JEL codes: C23.

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

[*] We 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 CAM, CEMFI, Columbia University, CREST, EUI, Harvard/MIT, IFS, Princeton University, Sciences-Po, Tinbergen Institute, Toulouse School of Economics, UCL, University of Chicago, Yale University, 2012 Panel Data Conference in Paris, and ESEM conference in Málaga for useful comments. Support from the European Research Council/ ERC grant agreement n^{0}263107 is gratefully acknowledged. All errors are our own.

1 Introduction

Unobserved heterogeneity is central to applied economics. There is ample evidence that workers and firms differ in many dimensions that are unobservable to the econometrician (Heckman, 2001). Cross-country analyses also show evidence of considerable heterogeneity (e.g., Durlauf *et al.*, 2001). In view of this prevalence, the use of flexible empirical approaches to model unobserved heterogeneity has been advocated in the literature (e.g., Browning and Carro, 2007). In practice, however, there is a trade-off between specifying rich patterns of 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 deal with the presence of unobserved heterogeneity in a panel data context.

A widely used approach in applied work is to model heterogeneous features as unit-specific, time-invariant fixed-effects. Fixed-effects approaches (FE) are conceptually attractive, as they allow for unrestricted correlation between unobserved effects and covariates. When one is interested in measuring the effect of one particular covariate, general fixed-effects endogeneity is thus taken care of in estimation. However, allowing for as many parameters as individual units comes at a cost. This lack of parsimony implies that estimates of common parameters are subject to an "incidental parameter" bias that may be substantial in finite samples (Nickel, 1981, Hahn and Newey, 2004). Moreover, the unit-specific fixed-effects are typically poorly estimated in short panels, often preventing the researcher to make sense of unobserved heterogeneity estimates.[1] In addition, FE may not be that flexible either. Indeed, although standard FE approaches model cross-sectional heterogeneity in an unrestricted way, they severely restrict its time-series variation by assuming that unobserved heterogeneity is constant over time.

This paper proposes a framework that allows for time patterns of unobserved heterogeneity that are common within groups of individuals. Both the group-specific time patterns and individual group membership are left unrestricted, and are estimated from the data. In particular, our time-varying specification shares with FE the fact that it leaves the relationship between observables and unobservables unrestricted, thus allowing 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 $g_i \in \{1, ..., G\}$ are

[1] As an example, estimation of the fixed effects has received some attention in the literature on school and teacher quality (Kane and Staiger, 2002). In short panels, it may be possible to consistently estimate features of the cross-sectional distribution of the individual fixed-effects, as opposed to the individual effects themselves (Arellano and Bonhomme, 2011).

unrestricted, and will be estimated along with the other parameters of the model. The group-specific time dummies α_{gt}, for $g \in \{1, .., G\}$, are fully unrestricted as well. As an example, all units in the first group (that is, all i such that $g_i = 1$) share the same unrestricted time profile α_{1t}. Lastly, the number of groups G is to be set or estimated by the researcher. The baseline framework of model (1) may easily be modified to incorporate restrictions on the group-specific time patterns, and to allow for additive time-invariant fixed-effects in addition to the time-varying grouped effects.

There are theoretical and empirical reasons for considering group-specific patterns of heterogeneity. As a first example, the static group interaction model for panel data (e.g., Blume *et al.*, 2010) may be seen as a special case of model (1), where α_{g_it} includes group means of covariates and outcomes. In this context, our framework may be used to estimate the reference groups, simply by treating α_{g_it} as unrestricted parameters. As a second example, tests of full risk-sharing in village economies are also often based on the same model (Townsend, 1994, Munshi and Rosensweig, 2009).[2] Note that, in contrast with most applications of social interactions and risk sharing models, our approach allows to estimate the reference groups from the data, under the assumption that group membership remains constant over time.

From an empirical perspective, in many applications interdependence across units is treated as a nuisance and taken into account using robust ("clustered") standard errors formulas. Yet, as illustrated by our empirical application on country-level data, dependence *per se* may often be of interest to the researcher. In this perspective, grouped patterns of heterogeneity can be interpreted as a flexible way of modelling interdependence across individual units over time. Compared to existing spatial dependence models for panel data (e.g., Sarafidis and Wansbeek, 2012), model (1) allows the researcher to estimate the spatial weights matrix. This relaxes an important requirement of these models, where the notion of "economic distance" is sometimes elusive.

A distinctive feature of our approach is that group membership is estimated from the data. 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. As the number of time periods increases, the accuracy of group classification improves. This approach is statistically well grounded, as it asymptotically recovers the population parameters when the model is correctly specified, and as it allows the researcher to compute standard errors that take into account the fact that groups have been estimated. Nonetheless, in contexts where the researcher has some *a priori* information on group composition, our estimator may easily be extended to incorporate this information to the estimation problem in a non-dogmatic way.

Determining the optimal grouping of the cross-sectional units represents a computational chal-

[2] In tests of full insurance, y_{it} in model (1) would be (the first difference of) log household consumption. An important assumption for the test to be valid is that households have common (CRRA) preferences, see for example Schulhofer-Wohl (2011).

lenge. We take advantage of the fact that this problem has been extensively studied by the research community working on data clustering (Steinley, 2006). In the absence of covariates in model (1), the estimation problem coincides with the standard minimum sum-of-squares partitioning problem, and a simple solution is given by the "kmeans" algorithm (Forgy, 1965). Making use of the connection with the clustering literature, we compute the GFE estimator using a state-of-the-art heuristic approach (Hansen *et al.*, 2010), which we extend to allow for covariates. Our algorithm delivers a fast and reliable answer to the computation problem.[3] We assess its performance by building on recently proposed *exact* solution algorithms (Brusco, 2006, Aloise *et al.*, 2009). The numerical experiments that we have performed suggest that our algorithm correctly identifies the globally optimal grouping, at least in datasets of moderate size such as the one that we use in our empirical application. This encouraging evidence is consistent with previous results obtained for minimum sum-of-squares partitioning (Brusco and Steinley, 2007).

We derive the statistical properties of the grouped fixed-effects estimator in an asymptotic where N (the number of units) and T (the number of time periods) tend to infinity simultaneously. Although the estimator is biased for small T, the bias vanishes at a faster-than-polynomial rate provided groups are well separated, and errors v_{it} satisfy suitable tail and dependence conditions. Under these assumptions, the GFE estimator is thus automatically higher-order bias-reducing. Our large-N, T analysis provides a formal justification for clustering methods, and complements existing results on minimum sum-of-squares partitioning (Bryant and Williamson, 1978, Pollard, 1981).

As the two dimensions of the panel diverge, the GFE estimator is asymptotically equivalent to the infeasible least squares estimator with known population groups. This finding implies that, 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. Building on Pollard (1982)'s analysis of the minimum sum-of-squares partitioning problem, we propose two methods to estimate the fixed-T variance. In a Monte Carlo exercise calibrated to the empirical application, we provide evidence that using either of these two methods in combination with the GFE estimator 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 well-documented positive association between income and democracy disappears when controlling for additive country- and time-effects in a panel dataset. They interpret the country-specific fixed-effects as reflecting long-run, historical factors that have shaped political and economic development of countries.

We revisit the evidence using the grouped fixed-effects estimator. Our benchmark results are based

[3] Stata codes, which make use of a Fortran executable program, are available at: http://www.cemfi.es/∼bonhomme/

on model (1), which allows for time-varying grouped patterns of unobserved country heterogeneity. There are several reasons to view GFE estimates as a useful complement to the FE results in 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. In addition, the GFE estimator is well suited to deal with the fact that the within-country variance of income is small, and the estimates of country-specific heterogeneity are imprecise due to the short length of the panel– seven five-year periods in our benchmark dataset (1970-2000). Lastly, the modelling of grouped patterns is also consistent with the empirical observation that regime types and transitions tend to cluster in time and in space, as documented in the political science literature (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.

According to the baseline specification, the effect of income on democracy remains positive and significant when allowing for grouped patterns of heterogeneity, although this effect is quantitatively small as a result of a substantial endogeneity bias in the cross-section. Moreover, the income effect disappears when allowing for time-varying grouped effects and time-invariant country-specific fixed-effects simultaneously. Our main empirical finding is that estimates of the time-varying unobserved determinants of democracy are only partly consistent with an additive fixed-effects specification. Specifically, while the unobserved determinants of democracy in two thirds of the sample countries display stable time profiles, those in the remaining third experience marked upward trends over the period. In our preferred specification, two of the four groups that we identify comprise transition countries– mostly located in Southern Europe and Latin America, and in part of Africa– whose democracy levels show substantial increases starting at different points in time.

An important question is then why the estimated time profiles differ across countries. To explore this issue, we relate the estimated groups to various factors that the literature has pointed out as potential determinants of democracy. In particular, we use a measure of constraints on the executive at the time of independence constructed by Acemoglu *et al.* (2005, 2008) as a historical, long-run determinant. We find that constraints at independence were significantly more stringent in countries that remained democratic between 1970 and 2000, compared to those that remained non-democratic during the period. However, this measure does not explain why some countries that were non-democratic at the beginning of the sample period experienced a democratic transition, while others did not. These results call for further study of the short and long-run determinants of democracy: for a sizable share of the world, history appears to have evolved at a fast pace.

Literature review and outline. The grouped fixed-effects estimator relates to the recent panel data literature. Its asymptotic properties contrast with available results for models with unit-specific fixed effects, where the incidental parameter bias is of the $O(1/T)$ order in general (Arellano and

Hahn, 2007). At the heart of this difference is the fact that group classification improves very fast as the number of time periods increases. A related result was recently obtained by Hahn and Moon (2010) in a class of nonlinear models with discrete time-invariant heterogeneity. Relative to Hahn and Moon, this paper allows for time-varying heterogeneity and provides primitive conditions in the case of the linear model, covering the minimum sum-of-squares partitioning problem as a special case. Interestingly, we also find that adding non-dogmatic prior information to GFE does not affect the large-T properties of the estimator, in contrast with fixed-effects models (Arellano and Bonhomme, 2009).

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 observed covariates.[4] 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 the unobserved heterogeneity, while leaving other features of the relationship between observables and unobservables unrestricted.[5]

The grouped fixed-effects estimator is also related to factor-analytic, "interactive fixed-effects" estimators (Bai, 2009). Indeed, the GFE model of unobserved heterogeneity has a factor-analytic structure, as:

$$\alpha_{g_i t} = \underbrace{(\alpha_{1t}, \alpha_{2t}, ..., \alpha_{Gt})}_{f_t'} \times \underbrace{(0, 0, ..., 1, ..., 0)'}_{\lambda_i}.$$

Unlike interactive FE, which recover the structure of heterogeneity up to an unknown rotation, the GFE approach recovers the exact group structure. In addition, for a given number of groups the GFE approach is more parsimonious than factor-analytic ones, resulting in smaller asymptotic biases under correct specification. This parsimony may be useful in situations where the data are not informative enough to allow for fully unrestricted interactive effects.

We take advantage of the mathematical connection with interactive fixed-effects models to conduct the asymptotic analysis. In particular, we use an insight from Bai (1994, 2009) to establish consistency of the GFE estimator. We also rely on the analysis of Moon and Weidner (2010b) to discuss the important issue of misspecification of the number of groups G. As an example, we show that estimating two groups when the data generating process is homogeneous does not bias the slope estimator.

[4] See the monographs by McLachlan and Peel (2000) and Frühwirth-Schnatter (2006) for recent advances in this area. Important contributions in economics include Heckman and Singer (1984) and Keane and Wolpin (1997). Kasahara and Shimotsu (2009) and Browning and Carro (2011) study identification in finite mixtures of discrete choice models for a fixed number of groups. Geweke and Keane (2007) and Norets (2010) are recent examples of flexible modelling strategies.

[5] In this sense, our approach is reminiscent of sparsity assumptions that have been widely studied in regression models (Tibshirani, 1996).

However, the bias on the intercept(s) can be substantial. Lastly, we rely on Bai and Ng (2002) to propose a class of information criteria that consistently select the true number of groups as N and T tend to infinity.

Finally, note that this paper is not the first one to rely on group structures for modelling unobserved heterogeneity in panels. Bester and Hansen (2010) show that grouping individual fixed-effects may result in gains in precision. In their setup, heterogeneity is time-invariant and the grouping of the data is assumed known to the researcher. A recent paper by Lin and Ng (2011) considers a random coefficients model and uses the time-series regression estimates to classify individual units into several groups. They also propose a classification algorithm that is related to ours, although they do not derive the asymptotic properties of the corresponding estimator. None of these two papers allows for time-varying unobserved heterogeneity.[6]

The outline of the paper is as follows. In Section 2 we introduce the grouped fixed-effects estimator and several extensions. In Section 3 we discuss computation issues. In Section 4 we derive the asymptotic properties of the estimator as N and T tend to infinity. Section 5 considers inference and estimation of the number of groups, and provides some finite sample evidence on the performance of the estimator. In Section 6 we use the GFE approach to study the relationship between income and democracy. Lastly, Section 7 concludes. Additional material may be found in a supplementary appendix.[7]

2 The grouped fixed-effects estimator

We start by introducing the grouped fixed-effects (GFE) estimator in the baseline model (1). Then we outline several extensions.

2.1 Baseline model

Model (1) contains three types of parameters: the parameter vector $\theta \in \Theta$, which is common across individual units; the group-specific time dummies $\alpha_{gt} \in \mathcal{A}$, 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.

It is assumed that x_{it} and v_{it} are weakly uncorrelated. In particular, the covariates vector x_{it} may include strictly exogenous regressors, lagged outcomes, or general predetermined regressors. The model

[6] 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). Yet another related work is Sun (2005), who considers parametric finite mixture models for panel data and studies the properties of maximum likelihood estimation.

[7] Attached at the end of this manuscript.

also allows for time-invariant regressors under certain support conditions. In contrast, x_{it} and $\alpha_{g_i t}$ are allowed to be arbitrarily correlated. We defer a more precise statement of the model assumptions until Section 4.

The grouped fixed-effects estimator is defined as the solution to 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 computational purposes, as well as to derive asymptotic properties, it is convenient to introduce an alternative characterization of the GFE estimator based on concentrated group membership variables. For given values of θ and α, the optimal 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) is then equivalently 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 $\widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right)$.

Two remarks are in order. First, unlike standard finite mixture modelling, which specifies the group probabilities as parametric or semiparametric functions of observed covariates (e.g., McLachlan and Peel, 2000), the grouped fixed-effects approach leaves group probabilities unrestricted. In fact, we show in the supplementary appendix 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 viewed as a point of contact between finite mixtures and fixed-effects.

Secondly, 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 $i = 1, ..., N$. However, the number of partitions of N units into G groups increases steeply with N, making exhaustive search virtually impossible. In the next section we will exploit recent advances in the literature on data clustering for efficient computation.

2.2 Extensions

Here we outline several simple extensions of the baseline grouped fixed-effects model that may be useful for applied work. We end the section by briefly describing a general GFE estimator for nonlinear models.

Unit-specific heterogeneity. One simple generalization is to allow for both time-invariant fixed-effects and time-varying grouped effects as follows:

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

where η_i are N unrestricted parameters. Denoting unit-specific means as $\overline{w}_i = \frac{1}{T}\sum_{t=1}^{T} w_{it}$, (5) yields the following equation in deviations to the mean:

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

which has the same structure as model (1) and may be estimated using grouped fixed-effects.[8]

Modelling time patterns. Another simple extension is to impose linear constraints on the group-specific time effects α_{gt}. For example, one may specify: $\alpha_{gt} = \sum_{r=1}^{R} b_{gr}\psi_r(t)$, where $\psi_1, ..., \psi_R$ are known functions, and where b_{gr} are scalar parameters to be estimated. Linear constraints are easy to embed within the computational and statistical framework of model (1), and allow to model a wide variety of patterns of unobserved heterogeneity. As an example, in the empirical application we show estimates of a model with two different layers of heterogeneity (time-varying and time-invariant) that takes the following form:

$$y_{it} = x_{it}'\theta + \alpha_{g_{i1}t} + \eta_{g_{i1},g_{i2}} + v_{it}, \tag{7}$$

where $(g_{i1}, g_{i2}) \in \{1, ..., G_1\} \times \{1, ..., G_2\}$ indicates joint group membership.[9]

Adding prior information. In certain applications researchers may want to incorporate prior information on the structure of unobserved heterogeneity. For example, in a cross-country application one could think that countries in the same continent are more likely to belong to the same group. In such situations, one possibility is to impose the group structure on the data by assumption, e.g. by controlling for continent dummies possibly interacted with time effects. Another approach is to use our grouped fixed-effects estimator, which leaves the groups unrestricted and recovers them endogenously. An intermediate possibility is to combine *a priori* information on group membership with data information, simply by adding a penalty term to the right-hand side of (2). See the supplementary appendix for details on this alternative approach.

[8] Our asymptotic results imply that GFE yields large-N, T consistent estimates of the model's parameters in (6) in the presence of strictly exogenous covariates or lagged outcomes. In contrast, GFE is generally inconsistent when covariates are endogenous– that is, when $\mathbb{E}(x_{it}v_{it}) \neq 0$– in the absence of instruments.

[9] This model may be interpreted as a restricted version of model (1) with $G = G_1 \times G_2$ groups, and with linear constraints on the group-specific time dummies. Indeed, letting $\mu_{g_1 g_2 t} = \alpha_{g_1 t} + \eta_{g_1,g_2}$ it is easy to see that the following $G_1(G_2-1)(T-1)$ linear constraints are satisfied:

$$\mu_{g_1 g_2 t} - \frac{1}{T}\sum_{s=1}^{T}\mu_{g_1 g_2 s} - \frac{1}{G_2}\sum_{h=1}^{G_2}\mu_{g_1 h t} + \frac{1}{G_2 T}\sum_{h=1}^{G_2}\sum_{s=1}^{T}\mu_{g_1 h s} = 0, \quad \text{for all } (g_1, g_2, t)\,.$$

Nonlinear models. To conclude this section, we note that 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:

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} m_{it}\left(\theta, \alpha_{g_i t}\right). \tag{8}$$

This framework covers random coefficients models and likelihood models as special cases.[10] In particular, it encompasses static and dynamic discrete choice models. However, studying the properties of GFE in nonlinear models raises a number of challenges, which we do not address in this paper.

3 Computation

Computation of the grouped-fixed effects estimator is challenging due to the piecewise-quadratic nature of the criterion. Given its accused non-convexity, and its large number of local minima, direct minimization is not well-suited. As an alternative, we exploit a connection with data clustering and take advantage of recent developments in this literature in order to obtain fast and efficient computation methods.

3.1 Algorithms

We present two computation algorithms in turn: a simple iterative scheme, and a more efficient alternative.

A simple iterative algorithm. A simple strategy to minimize (4) is to iterate back and forth between group classification (computation of g_i) and estimation of the other parameters (θ and α), until numerical convergence. This may be done as in the following iterative algorithm:

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* $i \in \{1, ..., N\}$*:*

$$g_i^{(s+1)} = \underset{g\in\{1,...,G\}}{\operatorname{argmin}} \sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta^{(s)} - \alpha_{gt}^{(s)}\right)^2. \tag{9}$$

[10] A GFE estimator in the random coefficients model is obtained by taking $m_{it}\left(\alpha_{g_i t}\right) = \left(y_{it} - x_{it}'\alpha_{g_i t}\right)^2$. Note that in this case $\mathcal{A}$ is a subset of $\mathbb{R}^K$, where $K = \dim x_{it}$. A GFE estimator in a likelihood setup is obtained by taking $m_{it}\left(\theta, \alpha_{g_i t}\right) = -\ln f\left(y_{it}|x_{it}; \theta, \alpha_{g_i t}\right)$, where $f(\cdot)$ denotes a parametric density function.

3. Compute:

$$\left(\theta^{(s+1)}, \alpha^{(s+1)}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{N}\left(y_{it} - x_{it}'\theta - \alpha_{g_i^{(s+1)}t}\right)^2. \tag{10}$$

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 $g_i^{(s+1)}$ whose vector of time effects is closest (in an Euclidean sense) to her vector of residuals $y_{it} - x_{it}'\theta^{(s)}$. In the "update" step, θ and α are computed given the group assignment. Note that (10) corresponds to an OLS regression that controls for interactions of group indicators and time dummies.[11]

This iterative scheme is a clustering algorithm. Indeed, it coincides with the well-known *kmeans* algorithm (Forgy, 1965) in the special case where there are no covariates in the model (i.e., when $\theta = 0$).[12] In this case, (4) boils down to the standard minimum sum-of-squares partitioning problem:

$$\widehat{\alpha} = \underset{\alpha\in\mathcal{A}^{GT}}{\text{argmin}} \sum_{i=1}^{N}\left(\min_{g\in\{1,...,G\}}\sum_{t=1}^{T}(y_{it} - \alpha_{gt})^2\right). \tag{11}$$

In geometric terms, (11) amounts to finding a collection of "centers" α_1, α_2, ..., α_G in $\mathbb{R}^T$ such that the sum of the Euclidean distances between y_i and the closest center α_g is minimum. Due to its relevance in many different fields (such as astronomy, genetics or psychology), this problem has been extensively studied in operations research and computer science (Steinley, 2006).

It is easy to see that, in Algorithm 1, the objective function is non-increasing in the number of iterations. Numerical convergence is typically very fast. However, a drawback of Algorithm 1 is its dependence on the chosen starting values. One way to overcome this problem is to choose many random starting values, and then select the solution that yields the lowest objective. In the numerical experiments reported below we use the following method to generate starting values:[13]

1. Draw $\theta^{(0)}$ from some prespecified distribution supported on Θ.

2. Draw G units $i_1, i_2, ..., i_G$ in $\{1, ..., N\}$ at random, and set:

$$\alpha_{gt}^{(0)} = y_{i_g t} - x_{i_g t}'\theta^{(0)}, \qquad \text{for all } (g,t).$$

[11] 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.

[12] Note that similar iterative schemes will apply to more general (possibly nonlinear) models. See for example the literature on "clusterwise regression" in operations research (Späth, 1979, Caporossi and Hansen, 2005), and more recently Lin and Ng (2011).

[13] See Maitra, Peterson and Ghosh (2011) for a comparison of various initialization methods for the kmeans algorithm. Another simple initialization scheme that we have considered it to select $G + r$ units at random, and to set $\left(\theta^{(0)}, \alpha^{(0)}\right)$ as the global minimum of the GFE objective in that subsample. This can be done easily for low values of r. A practical advantage of this method is that the researcher does not need to prespecify a distribution for $\theta^{(0)}$. In our experiments, we observed little difference between the two initialization methods.

A more efficient algorithm. In practice, as in kmeans, a prohibitive number of starting values may be needed to obtain reliable solutions. The Variable Neighborhood Search method has been pointed out as the state-of-the-art heuristic to solve the minimum sum-of-squares partitioning problem (Hansen and Mladenović, 2001, Hansen *et al.*, 2010). We extend the specific algorithm used in Pacheco and Valencia (2003) and Brusco and Steinley (2007) to allow for covariates. The algorithm works as follows, where as before $\gamma = \{g_1, ..., g_N\}$ is a generic notation for a partition of the N units into G groups:

Algorithm 2 *(Variable Neighborhood Search)*

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

   Perform one assignment step of Algorithm 1 and obtain an initial grouping γ_{init}.

   Set $iter_{max}$ *and* $neigh_{max}$ *to some desired values.*

   Set $j = 0$.

   Set $\gamma^* = \gamma_{init}$.

2. *Set* n *to 1.*

3. *(Neighborhood jump) Relocate* n *randomly selected units to* n *randomly selected groups, and obtain a new grouping* $\gamma^{'}$.

   Perform one update step of Algorithm 1 and obtain new parameter values $\left(\theta^{'}, \alpha^{'}\right)$.

4. *Set* $\left(\theta^{(0)}, \alpha^{(0)}\right) = \left(\theta^{'}, \alpha^{'}\right)$, *and apply Algorithm 1.*

5. *(Local search) Starting from the grouping* $\gamma = \{g_1, ..., g_N\}$ *obtained in Step 4, systematically check all re-assignments of units* $i \in \{1, ..., N\}$ *to groups* $g \in \{1, ..., G\}$ *(for* $g \neq g_i$*), updating* g_i *when the objective function decreases; stop when no further re-assignment improves the objective function.*

   Let the resulting grouping be $\gamma^{''}$.

6. *If the objective function using* $\gamma^{''}$ *improves relative to the one using* γ^*, *then set* $\gamma^* = \gamma^{''}$ *and go to Step 2; otherwise, set* $n = n + 1$ *and go to Step 7.*

7. *If* $n \leq neigh_{max}$, *then go to Step 3; otherwise go to Step 8.*

8. *Set* $j = j + 1$. *If* $j > iter_{max}$, *then Stop; otherwise go to Step 2.*

Algorithm 2 combines two different search technologies. First, a local search (Step 5) guarantees that a local optimum is attained, in the sense that the solution cannot be improved by re-assigning any single individual to a different group. Note that solutions of Algorithm 1 do not necessarily correspond to local minima in this sense. Secondly, re-assigning several randomly selected units into randomly selected groups (Step 3) allows for further exploration of the objective function. This is done by means of neighborhood jumps of increasing size, where the maximum size of the neighborhood $neigh_{max}$ is chosen by the researcher. Local search allows to get around local minima that are close to each other, whereas random jumps aim at efficiently exploring the objective function while avoiding to get trapped in a valley.

Algorithm 2 depends on two parameters set by the researcher: the maximum neighborhood size $neigh_{max}$, and a maximum number of iterations $iter_{max}$. The algorithm may also be run using different starting parameter values, even though the choice of starting values tends to matter much less than in the case of Algorithm 1. Denoting as N_s the number of starting values, Algorithm 2 is thus indexed by $(N_s; neigh_{max}; iter_{max})$.

3.2 Numerical performance

Tables 1 and 2 show the value of the final objective corresponding to different computation methods, on the cross-country panel dataset that we use in the empirical application. The dataset is described in Section 6, but for now it is enough to keep in mind its dimensions: $N = 90$, $T = 7$, and two covariates (including a lagged outcome). We show the value of the objective as well as computation time for both algorithms, and for $G = 2$, 3, and 10. In addition, we show the results for the first 30 countries, the first 60 countries (alphabetically ordered), and all 90 countries in the dataset, respectively.

Table 1 suggests that the simple iterative algorithm performs well when the number of groups is small. Algorithms 1 and 2 yield the same solution (that is, the same objective and optimal grouping) in all configurations of the data. In contrast, Table 2 shows that Algorithm 2 improves on Algorithm 1 when the number of groups increases. When $G = 10$ and $N = 30$, running the iterative algorithm using $1,000$ starting values yields a higher value for the objective function than when using Algorithm 2. When all $N = 90$ countries are included in Table 2, even $1,000,000$ different starting values and a running time of approximately one hour yields a higher objective than when using Algorithm 2 during only four minutes of search (7.749 versus 7.762, respectively).[14] Interestingly, running Algorithm 2 during 36 hours yields exactly the same objective and grouping.

Despite these results, one concern is that even the best heuristic methods can lead to non-optimal solutions. To assess whether the solutions of Algorithm 2 are optimal in Tables 1 and 2, we make use of– and extend– *exact* solution algorithms for the minimum sum-of-squares partitioning problem. New methods have recently been proposed to compute globally optimal solutions in this challenging problem,[15] including Brusco (2006)'s repetitive branch and bound algorithm, and Aloise *et al.* (2009)'s column generation algorithm. In the "exact" columns of Tables 1 and 2 (indicated with two or three stars) we report the objective function obtained when applying one of these exact algorithms to the

[14] The computer used in our calculations has 64 bits and 24 GB RAM.

[15] It has been proved that problem (11) may be solved exactly in $O(N^{GT+1})$ operations (Inaba *et al.*, 1994).

Table 1: Numerical performance ($G = 2, 3$)

$G = 2$

	Alg. 1 (1,000)		Alg. 2 (10;10;10)		Exact
	Value	time	Value	time	Value
$N = 30$	6.159	.6	6.159	2.1	6.159*
$N = 60$	13.209	.9	13.209	7.6	13.209*
$N = 90$	19.846	1.3	19.846	18.2	19.846*

$G = 3$

	Alg. 1 (1,000)		Alg. 2 (10;10;10)		Exact
	Value	time	Value	time	Value
$N = 30$	4.913	.6	4.913	6.1	4.913*
$N = 60$	10.934	1.1	10.934	16.7	10.934**
$N = 90$	16.598	1.7	16.598	38.4	16.598**

*Note: Balanced panel dataset from Acemoglu et al. (2008), $T = 7$, two covariates. Results for Algorithm 1 (N_s), with N_s randomly chosen starting values; and for Algorithm 2 ($N_s; neigh_{max}; iter_{max}$), with N_s starting values, maximum size of neighborhoods $neigh_{max}$, and maximum number of iterations $iter_{max}$. The value of the final objective and CPU time (in seconds) are indicated. In the "exact" column, ** refers to Brusco (2006)'s exact branch and bound algorithm for given $\widehat{\theta}$, and * refers to our extension of Brusco's algorithm that allows for covariates.*

Table 2: Numerical performance ($G = 10$)

	Alg. 1 (1,000)		Alg. 1 (1,000,000)		Alg. 2 (10;10;10)		Alg. 2 (1,000;20;20)		Exact
	Value	time	Value	time	Value	time	Value	time	Value
$N = 30$	1.106	1.1	1.025	988.3	1.025	48.3	1.025	10872.2	1.025**
$N = 60$	4.373	2.0	4.255	1729.5	4.255	116.4	4.255	28301.9	N/A
$N = 90$	8.035	3.4	7.762	3235.6	7.749	228.4	7.749	132555.7	7.749***

*Note: See note to Table 1. In the "exact" column, *** refers to Aloise et al. (2009)'s exact column generation algorithm for given $\widehat{\theta}$.*

vector of residuals $y_{it} - x_{it}'\widehat{\theta}$, where $\widehat{\theta}$ has been computed using our best heuristic (Algorithm 2). We see that the objective and grouping coincide with the one identified by Algorithm 2 in all cases, including when $G = 10$. This provides very encouraging evidence on the performance of our algorithm, which confirms previous evidence obtained for minimum sum-of-squares partitioning (Brusco and Steinley, 2007).

In addition, we were able to extend Brusco (2006)'s repetitive branch and bound algorithm to allow for covariates.[16] Although our current implementation is limited to a small number of groups ($G = 2$ for $N \leq 90$, and $G = 3$ for $N = 30$) it yields the same solution as the one obtained using the heuristics; see the results indicated with one star in Table 1. This formally demonstrates that our heuristic algorithm has correctly identified the global minimum in these cases.

Overall, this section suggests that the computation problem for GFE is challenging, yet not impossible, thanks to recent advances in the clustering literature. Our main algorithm (Algorithm 2) delivers fast and reliable estimates, and we have provided evidence that the solutions obtained are globally optimal in the dataset of our empirical application. In larger datasets, the simple iterative algorithm (Algorithm 1) is a practical option.[17,18] Assessing the numerical performance of the two algorithms as the dimensions of the problem increase is a natural next step.

Finally, it is worth pointing out that research on exact computation algorithms is still in progress. Recent research for solving problem (11) has shown that sophisticated interior point methods can deliver exact solutions in competitive time in several large instances.[19] We view these approaches as a potentially useful complement to heuristic methods in order to compute the GFE estimator.

[16] Brusco's algorithm is available at: http://mailer.fsu.edu/~mbrusco/bbwcss.for. The extension of the algorithm that allows for covariates is available from the authors upon request.

[17] We have computed the GFE estimator on several other empirical datasets (not reported). In particular, we have compared our two algorithms using log-earnings data from the PSID, in a sample with $N = 2075$, $T = 26$, three covariates, and $G = 3$. Our experiments suggest that Algorithm 1 provides reliable (though not necessarily optimal) solutions on these data.

[18] In large datasets, an alternative is to proceed in three steps: first estimate the GFE estimator on a random subsample of size $n << N$, yielding $(\widehat{\theta}^{(0)}, \widehat{\alpha}^{(0)})$; then classify all N units in the entire sample based on $(\widehat{\theta}^{(0)}, \widehat{\alpha}^{(0)})$; finally estimate $(\widehat{\theta}, \widehat{\alpha})$ using an OLS regression on the estimated groups, using the entire sample. Though not numerically equal to the argument of the global minimum of the GFE objective function, this three-step estimator will be asymptotically equivalent to the latter in a large-N, T perspective, under the conditions spelled out in Section 4, provided $n \to \infty$. We thank Denis Chetverikov for pointing this out to us.

[19] While Brusco (2006)'s repetitive branch and bound algorithm computed the global minimum in (11) in Fisher's Iris data ($N = 150$, $T = 4$) for as much as $G = 10$ groups, Du Merle *et al.* (2001) and more recently Aloise *et al.* (2009) computed exact solutions in datasets of dimensions up to $N = 2310$ and $T = 19$, for $G = 250$ groups. Note that the algorithm of Aloise *et al.* (2009) that we used in Table 2 delivered the global optimum in 1.7 seconds only.

4 Large-N, T asymptotic properties

In this section we characterize the asymptotic properties of the grouped fixed-effects estimator as N and T tend to infinity simultaneously. We provide conditions under which estimated groups converge to their population counterparts, and the incidental parameter bias due to misclassification shrinks to zero at a faster-than-polynomial rate as T tends to infinity. This implies that, in a large-T perspective, the GFE estimator is asymptotically equivalent to an infeasible least-squares estimator, even when T diverges (polynomially) more slowly than N.

4.1 The setup

In this first part we set the framework and provide some intuition for the main results. We 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 next 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 the GFE estimator is asymptotically equivalent to the infeasible least-squares estimator $\left(\widetilde{\theta}, \widetilde{\alpha}\right)$ as N and T tend to infinity and, for some $\nu > 0$, $N/T^\nu \to 0$. In particular, this allows T to grow considerably more slowly than N (when $\nu \gg 1$). To show asymptotic equivalence we assume that groups are well-separated. The result does not hold uniformly with respect to the values of the group-specific parameters. In the next section we will study an example where group separation fails.

Intuition in a simple case. Before discussing the general case of model (12), we start by providing an intuition in a simple case. Let us 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 $G = G^0 = 2$. The model is thus:

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

We further assume that $\alpha_1^0 \neq \alpha_2^0$, and in particular that the distance between the two remains positive as the sample size increases. In addition here we take $\alpha_1^0 < \alpha_2^0$ without loss of generality.

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 an individual who belongs to group 1 into group 2 is:

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

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 the number of parameters $g_1^0, ..., g_N^0$ tends to infinity with N. As a result, $\widehat{\alpha}$ 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.

In this simple model, it can easily be shown that, for $g = 1, 2$, the difference between $\widehat{\alpha}_g$ and the infeasible sample mean

$$\widetilde{\alpha}_g = \frac{\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\} \overline{y}_i}{\sum_{i=1}^N \mathbf{1}\left\{g_i^0 = g\right\}}$$

vanishes at an exponential rate as T tends to infinity. Hence if $N/T^\nu \to 0$ for some $\nu > 0$, then $\sqrt{NT}\left(\widehat{\alpha}_g - \widetilde{\alpha}_g\right)$ tends to zero asymptotically and $\widehat{\alpha}$ and $\widetilde{\alpha}$ have the same asymptotic distribution. Note that this result is specific to models with *discrete* heterogeneity: when α_i can take continuous values, in contrast, biases due to the incidental parameter problem are typically of the $O(1/T)$ order, and asymptotic equivalence with an unbiased infeasible target only holds if $N/T \to 0$ (e.g., Nickel, 1981, Hahn and Newey, 2004).

Extending the analysis of model (14) to a more general setup raises two main challenges. 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.[20] Secondly, the argument leading to the exponential rate of convergence of the misclassification probability (15) relies on the assumption that errors are i.i.d. normal. In order to bound tail probabilities under more general conditions (e.g., non-normality), approximations based on a central limit theorem are not sufficient. The analysis that we present next addresses both challenges.

[20] 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.

4.2 Properties

4.2.1 Consistency

We start by showing consistency under the following assumptions.

Assumption 1 *Let $M > 0$ be some constant.*

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

d. *For all* $g \in \{1, ..., G\}$*:* $\left|\frac{1}{NT}\sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \mathbb{E}\left(v_{it}v_{is}\alpha_{gt}^0\alpha_{gs}^0\right)\right| \leq M$*.*

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

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

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

h. *Let* $\overline{x}_{g\wedge\widetilde{g},t}$ *denote the mean of* x_{it} *in the intersection of groups* $g_i^0 = g$*, and* $g_i = \widetilde{g}$*.*[21] *Let* $\widehat{\rho}$ *be the minimum eigenvalue of the following matrix, where the minimum is taken over all possible groupings* $\gamma = \{g_1, ..., g_N\}$*:*

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

Then $\underset{N,T\to\infty}{\operatorname{plim}}\ \widehat{\rho} = \rho > 0$*.*

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 (fourth-order) 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.g. These are conceptually similar to assumptions commonly made in the literature on large factor models (Stock and Watson, 2002, Bai and Ng, 2002). Note that Assumptions 1.d and 1.e allow α_{gt}^0 and x_{it} to be weakly exogenous. In particular, this allows for lagged outcomes and general predetermined regressors, for example when

[21] 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.

$\mathbb{E}\left(v_{it}|x_{it}, x_{i,t-1}, ..., v_{i,t-1}, v_{i,t-2}, ...\right) = 0$. Assumptions 1.d, 1.e and 1.g impose conditions on the time-series dependence of errors (as well as covariates and time effects), while Assumption 1.f restricts the amount of cross-sectional dependence. Note that the latter condition is satisfied in the special case where v_{it} are i.i.d. across units.

Note also that Assumption 1.e may still be satisfied when errors and covariates are correlated with each other. A relevant example for applied work is model (5), when estimated in deviations to unit-specific means so as to remove time-invariant unit fixed-effects. In this case the assumption allows for predetermined regressors; for example, it is satisfied if $x_{it} = y_{i,t-1}$ under the conditions spelled out in Alvarez and Arellano (2003). When covariates are endogenous and Assumption 1.e does not hold, however, GFE leads to inconsistent estimates in general.

Lastly, Assumption 1.h is reminiscent of full rank conditions in standard regression models. We require that x_{it} shows sufficient variation over time and across individuals.[22] As a special case, the condition will be satisfied if x_{it} is 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. For example, if x_{it} follows a non-degenerate Bernoulli distribution, i.i.d in both dimensions, then $(x_{i1}, ..., x_{iT})$ has 2^T points of support, which may well be larger than $G+1$. As another special case, it can be shown that Assumption 1.h holds when x_{it} is i.i.d. normal.[23] Note also that Assumption 1.h allows for time-invariant regressors, provided that their support is rich enough.

We have the following result, where for conciseness we denote as $\widehat{g}_i = \widehat{g}_i\left(\widehat{\theta}, \widehat{\alpha}\right)$ 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:*

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

and:

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

Proof. See Appendix A. ■

The consistency proof is complicated by the fact that the dimension of α diverges as T tends to infinity. This prevents the adoption of standard techniques (e.g., Newey and McFadden, 1994) to prove the result. Instead, we build on an insight from Bai (1994, 2009) and consider an auxiliary objective function whose minimum is attained at $\left(\theta^0, \alpha^0\right)$. The strategy of the proof consists then in showing that the difference between the GFE objective function and the auxiliary one becomes uniformly small as N and T tend to infinity.

[22] Assumption 1.h is interestingly related to Assumption A in Bai (2009).

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

4.2.2 Asymptotic equivalence

We make the following assumptions.

Assumption 2 *Let $a, b, d_1, d_2 > 0$ be constants.*

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

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

c. For all $i \in \{1, ..., N\}$ and all $g \in \{1, ..., G\}$, $\{v_{it}\}_t$ and $\{\alpha_{gt}^0\}_t$ are strongly mixing processes with mixing coefficients that satisfy $\alpha[t] \leq e^{-at^{d_1}}$ for all t.[24] Moreover, $\mathbb{E}\left(\alpha_{gt}^0 v_{it}\right) = 0$.

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

e. One of the two following conditions holds:

(i) x_{it} has bounded support in $\mathbb{R}^K$.

(ii) $\{\|x_{it}\|\}_t$ satisfies the mixing and tail conditions of 2.c and 2.d above.

In contrast with consistency, we restrict the analysis of the asymptotic distribution to the case where the G population groups are well-separated (Assumptions 2.a and 2.b). In general the properties of the GFE estimator are different if group separation fails, for example when the number of groups in the population is strictly smaller than the number of groups postulated by the researcher (i.e., when $G^0 < G$). In the next section we will 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} is strongly mixing with a faster-than-polynomial decay rate, with tails also decaying at a faster-than-polynomial rate. The process α_{gt}^0 is assumed to be strongly mixing as well, and to be contemporaneously uncorrelated with v_{it}. Note that this strengthens the assumptions made in Assumption 1 regarding time-series dependence. These conditions allow us to rely on exponential inequalities for dependent processes (e.g., Rio, 2000) in order to bound misclassification probabilities.[25]

Finally, in Assumption 2.e we impose either of two conditions on covariates x_{it}. In Part (i) we require that covariates have bounded support. Alternatively, in Part (ii) we require that covariates satisfy dependence and tail conditions similar to the ones on v_{it}. A relevant case where the strong

[24] 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}.

[25] It is possible to relax Assumptions 2.c-2.e and assume that v_{it}, α_{gt}^0, and possibly $\|x_{it}\|$, are strongly mixing with a polynomial decay rate, and that their marginal distributions have 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 $d_2 > 2$. Under these weaker assumptions, it may be shown that $\widehat{\theta} - \widetilde{\theta} = o_p\left(T^{-q}\right)$, provided that $\frac{(d_1+1)d_2}{d_1+d_2} > 4q + 1$.

mixing conditions of Assumption 2.e (ii) may not necessarily hold is 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. For this reason, the proof of Theorem 2 below also explicitly addresses this case.

The next result shows that the GFE estimator and the infeasible least squares estimator with known population groups 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:*

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

and:

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

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

Proof. See Appendix A. ■

It follows from Theorem 2 that the asymptotic distribution of the grouped fixed-effects estimator and that of the infeasible least squares estimator coincide if, for some $\nu > 0$, N/T^{ν} tends to zero as N and T tend to infinity simultaneously. For example we have, using (17):

$$\sqrt{NT}\left(\widehat{\theta} - \widetilde{\theta}\right) = o_p\left(N^{\frac{1}{2}}T^{\frac{1}{2}-\delta}\right),$$

which is $o_p(1)$ as soon as $\delta \geq (\nu + 1)/2$. In addition, under these relative rates of N and T the estimated groups are uniformly consistent for the population ones, in the sense that:

$$\sup_{i\in\{1,\ldots,N\}} \left|\widehat{g}_i - g_i^0\right| \xrightarrow{p} 0. \tag{19}$$

Note that these relative rates allow T to increase polynomially more slowly than N. In contrast, the large-N, T asymptotic analysis of FE estimators is typically done assuming that $N/T \to C^{st}$, or that $N/T^3 \to 0$ in the case of bias-reduced estimators (e.g., Arellano and Hahn, 2006). Asymptotic equivalence as $N/T^{\nu} \to 0$ is the consequence of the fact that, unlike most fixed-effects or interactive fixed-effects estimators, the GFE estimator is unbiased to *any* (polynomial) order of magnitude relative to the infeasible least-squares estimator.

4.2.3 Asymptotic distribution

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:

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

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. In particular, Assumption 3.a ensures that the estimator has no asymptotic bias. Note that this condition is satisfied if x_{it} is 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 $y_{i-1,t}$). Note however that this assumption rules out lagged outcomes in model (5) with additive fixed-effects. Indeed, in deviations to unit-specific means we have: $\mathbb{E}\left[\left(v_{it}-\overline{v}_i\right)\left(y_{i,t-1}-\overline{y}_{i,-1}\right)\right]\neq 0$, and the least-squares estimator of θ suffers from an $O(1/T)$ bias.[26] Lastly, Assumptions 3.d-3.e similarly ensure that $\widetilde{\alpha}_{gt}$ has a standard asymptotic distribution.

We have the following result.

[26] Note that the GFE estimates of group membership are consistent in model (5) if the conditions of Theorem 2 are satisfied in deviations to unit-specific means. In the presence of lagged outcomes in (5), one could thus estimate $g_1, ..., g_N$ using GFE, and in a second step estimate the other parameters using any dynamic panel data estimator conditional on the estimated group dummies and time dummies. The large-N, T properties of this type of two-step estimators would require a separate analysis, which exceeds the scope of this paper.

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:*

$$\sqrt{NT}\left(\widehat{\theta}-\theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\theta}^{-1}\Omega_{\theta}\Sigma_{\theta}^{-1}\right), \tag{20}$$

and, for all (g,t):

$$\sqrt{N}\left(\widehat{\alpha}_{gt}-\alpha_{gt}^0\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\pi_g^2}\right), \tag{21}$$

where π_g is defined in Assumption 2, and where Σ_{θ}, Ω_{θ}, and ω_{gt} are defined in Assumption 3.

Proof. See Appendix A. ■

We end this section with two remarks. Firstly, suppose one wants to fit a parametric model (e.g., an ordered probit or a multinomial logit model), indexed by a parameter vector ξ, to the estimated groups:

$$\widehat{\xi} = \underset{\xi}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{g=1}^{G} \mathbf{1}\left\{\widehat{g}_i = g\right\} \ln\left(p_g\left(x_i;\xi\right)\right),$$

where $p_g(x;\xi)$ are the parametrically specified group probabilities. Then, in a large-N,T perspective and under similar conditions as in Theorem 2, $\widehat{\xi}$ will be asymptotically equivalent to the following infeasible maximum likelihood estimator:

$$\widetilde{\xi} = \underset{\xi}{\operatorname{argmax}} \sum_{i=1}^{N}\sum_{g=1}^{G} \mathbf{1}\{g_i^0 = g\} \ln\left(p_g\left(x_i;\xi\right)\right).$$

This implies that parameter estimates (and their standard errors) that treat the estimated groups as data will be asymptotically valid.

Secondly, note that the equivalence result in Theorem 2 still holds when considering a penalized version of the grouped fixed-effects estimator that incorporates non-dogmatic prior information, as we show in the supplementary appendix. In FE models, adding prior information on the individual effects has generally a first-order effect on the bias of the estimator (Arellano and Bonhomme, 2009). In contrast, in models where unobserved heterogeneity is discrete, and under the conditions of Theorem 2, adding non-dogmatic prior information has no effect on the asymptotic distribution of the estimator as N and T tend to infinity and N/T^{ν} tends to zero.

5 Inference and choice of the number of groups

In this section we discuss two important practical issues: estimation of the covariance matrices and estimation of the number of groups. In addition, we show the results of a small simulation experiment aimed at assessing the finite sample performance of our estimator, as well as that of our inference methods and choice of the number of groups.

5.1 Estimating covariance matrices

Large-N, T inference. We start with estimation of the large-T variance of group-specific time effects and common parameters under the conditions of Corollary 1. Assuming independent observations across individual units,[27] the variance of $\widehat{\alpha}_{gt}$ for all g, t can be estimated using the White formula:

$$\widehat{\text{Var}}\left(\widehat{\alpha}_{gt}\right) = \frac{\sum_{i=1}^{N} \mathbf{1}\left\{\widehat{g}_i = g\right\} \widehat{v}_{it}^2}{\left(\sum_{i=1}^{N} \mathbf{1}\left\{\widehat{g}_i = g\right\}\right)^2}, \tag{22}$$

where $\widehat{v}_{it} = y_{it} - x_{it}'\widehat{\theta} - \widehat{\alpha}_{\widehat{g}_i t}$ are the estimated GFE residuals.

Following Corollary 1, we estimate the asymptotic variance of $\widehat{\theta}$ as follows:

$$\widehat{\text{Var}}\left(\widehat{\theta}\right) = \frac{\widehat{\Sigma}_\theta^{-1}\widehat{\Omega}_\theta\widehat{\Sigma}_\theta^{-1}}{NT}, \tag{23}$$

where, denoting as $\overline{x}_{gt}$ the mean of x_{it} in group $\widehat{g}_i = g$,[28] we take:

$$\widehat{\Sigma}_\theta \quad = \quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it} - \overline{x}_{\widehat{g}_i,t}\right)\left(x_{it} - \overline{x}_{\widehat{g}_i,t}\right)',$$

and where $\widehat{\Omega}_\theta$ is a consistent estimate of the matrix Ω_θ.

In the presence of serial correlation, one may use the truncated kernel method of Newey and West (1987) in order to construct an estimator $\widehat{\Omega}_\theta$, as in Bai (2003). Alternatively, one may use the following formula clustered at the individual level (Arellano, 1987):

$$\widehat{\Omega}_\theta \quad = \quad \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\widehat{v}_{it}\widehat{v}_{is}\left(x_{it} - \overline{x}_{\widehat{g}_i,t}\right)\left(x_{is} - \overline{x}_{\widehat{g}_i,s}\right)'.$$

The properties of Arellano (1987)'s formula in FE models as N and T tends to infinity are studied in Hansen (2007). Below we show numerical evidence on the finite sample performance of the estimator (23) of the variance of the GFE estimator.

Large-N, fixed-T inference. The large-N, T asymptotic analysis of Section 4 provides conditions under which estimation of group membership does not affect inference. This result does not hold in an asymptotic when T is kept fixed as N tends to infinity. As we show in the supplementary appendix, the fixed-T variance of the grouped fixed-effects estimator reflects the additional contribution of observations that are at the margin between two groups, so that an infinitesimal change in parameter

[27] Note that the assumptions of Corollary 1 allow for weak dependence in the cross-sectional dimension. However, the variance formulas (22)-(23) are generally invalid in that case. A robust alternative is to follow Bai and Ng (2006), and to construct a partial sample variance estimator based on a random sample of size $n << \min(N, T)$.

[28] That is: $\overline{x}_{gt} = \frac{\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i = g\}x_{it}}{\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i = g\}}$. Note that this differs from the mean covariate defined in Assumption 3, as here the mean is computed within an *estimated* group.

values may entail re-classifying these observations. The analytical variance formula that we derive extends previous results by Pollard (1981, 1982) to allow for covariates.

However, a fixed-T asymptotic analysis is not directly informative to perform valid inference for the population parameters of the grouped model (12). Indeed, for fixed T the GFE estimator $\left(\widehat{\theta},\widehat{\alpha}\right)$ is root-N consistent and asymptotically normal for a *pseudo-true* value $\left(\overline{\theta},\overline{\alpha}\right)$, which minimizes an expected within-group sum of squared residuals. This pseudo-true value does not coincide with the true parameter value in general, although the difference between the two vanishes as T increases.

Nevertheless, a practical possibility to account for the effect of the estimation of group membership 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 estimators: the first one is a semiparametric estimator of the analytical variance formula, and the second one is based on the bootstrap. In the Monte Carlo exercise below we compare the finite-sample performance of these two estimators with the large-T variance estimator (22)-(23), and provide evidence that the former methods lead to more reliable inference for the population parameters in this case.

5.2 Unknown number of groups

The asymptotic results of Section 4 were derived under the assumption that the true number of groups G^0 is known. Here we relax this assumption and let G be the (possibly incorrect) number of groups postulated by the researcher.

Incorrect number of groups: a simple case. Misspecification of the number of groups has different effects on common parameter estimates, depending on whether the postulated number of groups is above or below the true one.

When $G < G^0$, the GFE estimator $\widehat{\theta}$ is generally inconsistent for θ^0 if the unobserved effects are correlated with the observed covariates. The inconsistency arises because of omitted variable bias. In contrast, when $G > G^0$ common parameters $\widehat{\theta}$ remain consistent for θ^0 under the conditions of Theorem 1, since the proof of the theorem is unaffected in this case. However, the group-specific effects may suffer from a substantial small-T bias, as the following simple example illustrates.

Proposition 1 *Let us consider the model:*

$$y_{it} = x_{it}'\theta^0 + \alpha^0_{g_i^0} + v_{it}, \qquad v_{it} \sim iid\mathcal{N}(0,\sigma^2), \tag{24}$$

where the true number of groups is $G^0 = 1$, and where $\alpha^0 = \alpha^0_1$ denotes the true value of α.

Let $\left(\widehat{\theta},\widehat{\alpha}\right)$ be the GFE estimator of (θ^0,α^0) with $G = 2$ groups. Then, as T is kept fixed and N tends to infinity we have: $\widehat{\theta} \xrightarrow{p} \theta^0$, and $\widehat{\alpha}_g \xrightarrow{p} \alpha^0 \pm \sigma\sqrt{\frac{2}{\pi T}}$, for $g = 1, 2$.

Proof. See Appendix A. ■

In this example, the data generating process is homogeneous ($G^0 = 1$), but the researcher estimates two groups ($G = 2$). The proof of Proposition 1 shows that, asymptotically, the two estimated groups are solely based on random errors (depending on whether $\overline{v}_i \geq 0$). Given that the spurious groups are independent of covariates, their presence does not bias the GFE estimator of θ^0. In fact, allowing for a larger number of groups than the true one in GFE estimation may be thought of as including $(G - G^0)$ irrelevant regressors– uncorrelated with the covariates of interest– in a linear regression. A similar intuition applies to interactive fixed-effects models: Moon and Weidner (2010b) show that the asymptotic distribution of the interactive FE estimator with $G \geq G^0$ factors is identical to that of the estimator based on the correct number of factors. We conjecture that this result applies to the GFE estimator in model (1). However, a formal proof of this conjecture is beyond the scope of this paper.

In contrast with common parameters, although the group effects $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$ are both consistent for α^0 as T tends to infinity, they suffer from a bias of order $O(1/\sqrt{T})$ for small T, which is one order of magnitude *larger* than the usual $O(1/T)$ order in FE panel data models. The $\sigma\sqrt{\frac{2}{\pi T}}$ term in Proposition 1 is simply the mean of a truncated normal $(0, \sigma^2/T)$ (i.e., the mean of $\overline{v}_i$ truncated at zero).

Estimating the number of groups. To consistently estimate the number of groups G^0 we rely on the connection with the analysis of large factor models and interactive fixed-effects panel data models and consider the following class of information criteria:

$$I\left(G\right) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\widehat{\theta}^{(G)} - \widehat{\alpha}_{\widehat{g}_i t}^{(G)}\right)^2 + Gh_{NT}, \tag{25}$$

where the $^{(G)}$ superscript refers to the grouped fixed-effects estimator with G groups. The estimated number of groups is then:

$$\widehat{G} = \underset{G\in\{1,\ldots,G_{max}\}}{\text{argmin}} \quad I\left(G\right), \tag{26}$$

where G_{max} is an upper bound on G^0, which is assumed known in order to derive the asymptotic properties.

Following the arguments in Bai and Ng (2002) and Bai (2009), it can be shown that the estimated number of groups $\widehat{G}$ is consistent for G^0 if, as N and T tend to infinity, h_{NT} tends to zero and $\min(N, T)h_{NT}$ tends to infinity. The first condition ensures that $\widehat{G} \geq G^0$ with probability approaching one, while the second condition guarantees that $\widehat{G} \leq G^0$.

As an example, let us consider the following Bayesian Information Criterion (BIC):[29]

[29] Given that unobserved heterogeneity is discrete, there is some ambiguity on how to define the number of parameters in the grouped fixed-effects approach. In (27) we have simply added the number of group-specific time effects (that is, GT), the number of common parameters (K), and the number of group membership variables g_i (that is, N). In the supplementary appendix we report simulation results using (27), as well as using an alternative choice with a steeper penalty.

$$BIC\left(G\right) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}^{\prime}\widehat{\theta}^{(G)}-\widehat{\alpha}_{\widehat{g}_it}^{(G)}\right)^2+\widehat{\sigma}^2\frac{GT+N+K}{NT}\ln(NT), \qquad (27)$$

where $\widehat{\sigma}^2$ is a consistent estimate of the variance of v_{it}.[30] One easily sees that the BIC estimate $\widehat{G}$ provides an upper bound on G^0 asymptotically if $(\ln T)/N \to 0$. In addition, $\widehat{G}$ is consistent for G^0 if N and T tend to infinity at the same rate. In contrast, if T tends to infinity more slowly than N so that T/N tends to zero, the BIC criterion (27) provides a conservative, possibly inconsistent, estimate of G^0.

5.3 A small-sample exercise

In the last part of this section, we study the suitability of our asymptotic results as a guide for small sample inference. We do this by means of a Monte Carlo exercise on simulated data, which we design to mimic the cross-country dataset that we will use in the empirical application.

Specifically, we consider the same sample size: $N = 90$ units and $T = 7$ periods. For a given number of groups, the data generating process follows model (12) where $x_{it} = (y_{i,t-1}, \widetilde{x}_{it})$ contains a lagged outcome and a strictly exogenous regressor, and where the process $\widetilde{x}_{it}$ is taken from the log-income per capita data. For this specification, we first estimate the model on the empirical dataset using grouped fixed-effects. Then, we fix the parameters of the DGP: θ^0, α^0 and all the group membership variables g_i^0, to their estimated GFE values. Lastly, the error terms are generated as i.i.d. normal draws across units and periods with variance equal to the mean of squared GFE residuals.

We start by showing the mean of the GFE estimator across $1,000$ Monte Carlo simulations in Table 3. We show the results for the two coefficients (θ_1 and θ_2, respectively), as well as for the "long-run" coefficient of $\widetilde{x}_{it}$ (i.e., $\theta_2/(1-\theta_1)$). Biases appear moderate despite the short length of the panel, at most 10% in relative terms. The last column in Table 3 shows the average misclassification frequency across simulations.[31] When $G = 3$ or 5, units are well classified in approximately 90% of cases. When $G = 10$, however, the frequency of correct classification drops to 55%. Nonetheless, the bias of the GFE estimator remains rather low. This suggests that the GFE estimator of common parameters may behave well in situations when G is not small relative to the sample size. In the conclusion, we shall

[30] A possibility is to estimate $\widehat{\theta}$, $\widehat{\alpha}$, and $\{\widehat{g}_1, ..., \widehat{g}_N\}$ using grouped fixed-effects with G_{max} groups, and to compute:

$$\widehat{\sigma}^2 = \frac{1}{NT-G_{max}T-N-K}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}^{\prime}\widehat{\theta}-\widehat{\alpha}_{\widehat{g}_it}\right)^2.$$

[31] The misclassification frequency is computed as $\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i \neq g_i^0\}$. To deal with invariance to relabelling we take, in each simulated sample, the labelling that yields the minimum amount of misclassification across all $G!$ permutations of group indices. When $G = 10$ this computation results prohibitive, so we take the minimum over $500,000$ randomly generated permutations.

comment on the possibility to modify the asymptotic analysis in order to allow G to increase together with N and T at a suitable rate.

Table 3: Bias of the GFE estimator

	θ_1 (coeff. $y_{i,t-1}$)		θ_2 (coeff. $\widetilde{x}_{it}$)		$\frac{\theta_2}{1-\theta_1}$		Misclassified
	True	GFE	True	GFE	True	GFE	
$G = 3$	.407	.391	.089	.099	.151	.163	9.50%
$G = 5$	.255	.262	.079	.086	.107	.117	9.68%
$G = 10$	.277	.286	.075	.078	.104	.110	44.73%

Note: The columns labelled "GFE" refer to the mean of GFE parameter estimates across $1,000$ *simulations. Algorithm 2– with parameters* $(5; 10; 5)$– *was used for computation. The last column shows the average of the misclassification frequency* $(\widehat{g}_i \neq g_i^0)$ *across simulations. Errors are i.i.d. normal.*

We next turn to inference. The top panel in Table 4 reports the standard deviation of the GFE estimator of θ across Monte Carlo simulations, together with the medians across simulations of three different standard errors estimates: the (square root of the) clustered variance formula (23), estimates based on Pollard (1982)'s fixed-T formula, and estimates based on the bootstrap.[32] The results show that the clustered formula systematically underpredicts the variability of the GFE estimator. This shows that group misclassification may have a sizable effect on inference in small samples. In contrast, the two fixed-T consistent estimates of the variance are larger, and more in line with the finite-sample dispersion. Moreover, the bottom panel in the table shows that these two methods provide approximately correct coverage for the true parameter θ^0, while estimates based on the large-T approximation tend to lead to overrejection.

Finally, in the supplementary appendix we show the results of several additional exercises. We first consider a natural alternative estimator, the interactive fixed-effects estimator of Bai (2009) with 3 factors, when the DGP follows the GFE model with $G^0 = 3$ groups. Although the interactive FE estimator is consistent as N and T tend to infinity, our results show that it suffers from a very substantial finite sample bias, much larger than the bias of the GFE estimator on this (relatively small) sample. Secondly, given that the asymptotic behavior of the GFE estimator crucially depends on tail and dependence properties of errors, we estimate a non-normal specification with dependent errors, finding similar results as in the main exercise.[33] We also report results for the group-specific time effects. Lastly, we provide evidence on the accuracy of the BIC criterion (27) to estimate the number of groups on the simulated data.

[32] Means of standard errors across simulations are very similar.

[33] This exercise is partly motivated by the fact that the measures of democracy that we use in the empirical application (Freedom House and Polity indices) take a small number of values.

Table 4: Inference for the GFE estimator

Standard errors

	θ_1 (coeff. $y_{i,t-1}$)				θ_2 (coeff. $\widetilde{x}_{it}$)				$\frac{\theta_2}{1-\theta_1}$			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$G = 3$	.035	.051	.068	.043	.0093	.0132	.0156	.0137	.013	.022	.030	.021
$G = 5$	.037	.068	.097	.058	.0088	.0135	.0160	.0112	.011	.022	.035	.022
$G = 10$	.037	.048	.091	.059	.0074	.0095	.0156	.0103	.009	.012	.026	.015

Coverage (nominal level 5%)

	θ_1 (coeff. $y_{i,t-1}$)			θ_2 (coeff. $\widetilde{x}_{it}$)			$\frac{\theta_2}{1-\theta_1}$		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
$G = 3$	.847	.965	.970	.723	.883	.914	.693	.917	.900
$G = 5$	.848	.961	.973	.788	.932	.943	.710	.936	.928
$G = 10$	.798	.902	.992	.841	.912	.996	.783	.904	.986

Note: Median standard errors across 1,000 *simulations (top panel) and empirical non-rejection probabilities (bottom panel, nominal size 5%). Column* (1) *reports results based on the large-T clustered variance formula (23),* (2) *reports estimates based on Pollard (1982)'s fixed-T formula, and* (3) *shows results based on the bootstrap (*100 *replications, Algorithm* 1 *with* 1,000 *starting values). Finally, column* (4) *in the top panel shows Monte Carlo standard deviations across simulations.*

6 Application: income and (waves of) democracy

In this last part of the paper we use the grouped fixed-effects approach to study the relationship between income and democracy across countries during the last third of the past century.

6.1 The empirical setup

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 to account for factors that simultaneously affect both economic and political development. Using country-level panel data, they document that the positive effect of income on democracy disappears when including country-specific fixed-effects in the regression. Acemoglu *et al.* (2008) argue that these results are consistent with countries having embarked on divergent paths of economic and political development at certain points in time, or *critical junctures*. Some of the examples they mention are the end of feudalism, the industrialization age, or the process of colonization (Acemoglu, Johnson and Robinson, 2001). In this perspective, the inclusion of fixed-effects in the regression is meant to capture these highly persistent long-run historical effects.

Table B1 in Appendix B replicates the main specification from Acemoglu *et al.*: a fixed-effects regression of democracy on lagged income per capita, using lagged democracy and time dummies as controls:[34]

$$democracy_{it} = \theta_1 democracy_{it-1} + \theta_2 logGDPpc_{it-1} + \alpha_{it} + v_{it}, \tag{28}$$

where $\alpha_{it} = \eta_i + \delta_t$. Democracy is measured according to the Freedom House indicator, and log-GDP per capita is taken from the Penn World tables. All data are taken at the five-year frequency. We consider two different samples: a balanced panel, which covers 90 countries on the period $1970-2000$, and an unbalanced panel, which covers 150 countries on the period $1960-2000$. According to the pooled OLS regressions, and regardless of the sample used, there is a statistically significant association between income and democracy. The point estimates of the cumulative income effect $\theta_2/(1-\theta_1)$ imply that a 10% increase in income per capita is associated with a 2.5% increase in the Freedom House score.[35] However, in both datasets, the FE estimates of the income coefficient are small or negative, and insignificant from zero.

In this section, we revisit the evidence using the grouped fixed-effects approach. Our main estimation results correspond to the baseline GFE specification: $\alpha_{it} = \alpha_{g_it}$, in which we allow for unrestricted group-specific time patterns of heterogeneity for several values of the number of groups G. In addition, we consider two other specifications: one that combines group-specific time-varying heterogeneity with country-specific time-invariant effects, that is $\alpha_{it} = \alpha_{g_it} + \eta_i$; and another one that contains two different layers of grouped heterogeneity: $\alpha_{it} = \alpha_{g_{i1}t} + \eta_{g_{i1},g_{i2}}$. These various specifications provide novel insights regarding the unobserved determinants of democracy and their evolution over time.

There are several reasons to think that the grouped fixed-effects approach provides a useful complement to FE in order to study the observed and unobserved determinants of democracy. First, a large literature in political science emphasizes time-varying determinants of political regimes (e.g., Przeworski *et al.*, 2000), which motivates the need to allow for time-varying unobserved heterogeneity. Second, the parsimonious nature of GFE is well-suited to deal with the small within-country variance of income (6% of the total variance of income in the balanced sample), and the short length of the panel (7 to 9 periods). Lastly, allowing for time-varying group-specific patterns of heterogeneity in this context is empirically motivated by the strong evidence of clustering of regime types and transitions, across time and space, documented in the political science literature (e.g., Gleditsch and Ward, 2006, Ahlquist and Wibbels, 2012).

A conceptual motivation for the grouped fixed-effects model can be found in Samuel Huntington's influential work on the "third wave" of democratization. Huntington (1991) emphasizes the

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

[35] To assess the magnitude of this effect, note that the Freedom House measure is normalized to lie between zero and one, and that its mean and standard deviation in the balanced sample are .55 and .37, respectively.

importance of international and regional factors as drivers of transitions to democracy and autocracy, resulting in groups of countries making transitions at similar points in time; that is, in "waves" of democratization.[36] Along with other examples, he mentions the influence of the US administration in the 1970s and changes in the Soviet Union in the early 1980s, the influence of the European Union in the late 1970s, or changes in the Catholic Church following the second Vatican council, as possible drivers of the clustering of transitions towards democracy that occurred between 1974 and 1990.

Huntington's arguments are consistent with the grouped fixed-effects model: for example $g_i = g$ could denote being predominantly Catholic, and α_{gt} could be the effect of the influence of the Catholic Church on the political evolution of the country. However, our estimation framework is agnostic about the causes of the "waves" of democracy, as it recovers heterogeneous patterns of political evolution from the data. In particular, we see our framework as a natural starting point to assess how well different theories of democratization fit the political and economic evolution of countries over time.

6.2 Common parameters: income and lagged democracy

We start by presenting the estimates of the coefficients of income and lagged democracy for the baseline grouped fixed-effects model. We report the results for the balanced subsample. Results for the unbalanced sample are qualitatively similar, and are summarized in Section 6.4 below.

Figure 1 plots the point-estimates and standard errors of income and democracy coefficients for different values of the number of groups G.[37] 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.[38]

According to our estimates, the cumulative income effect is statistically significant.[39] However, it

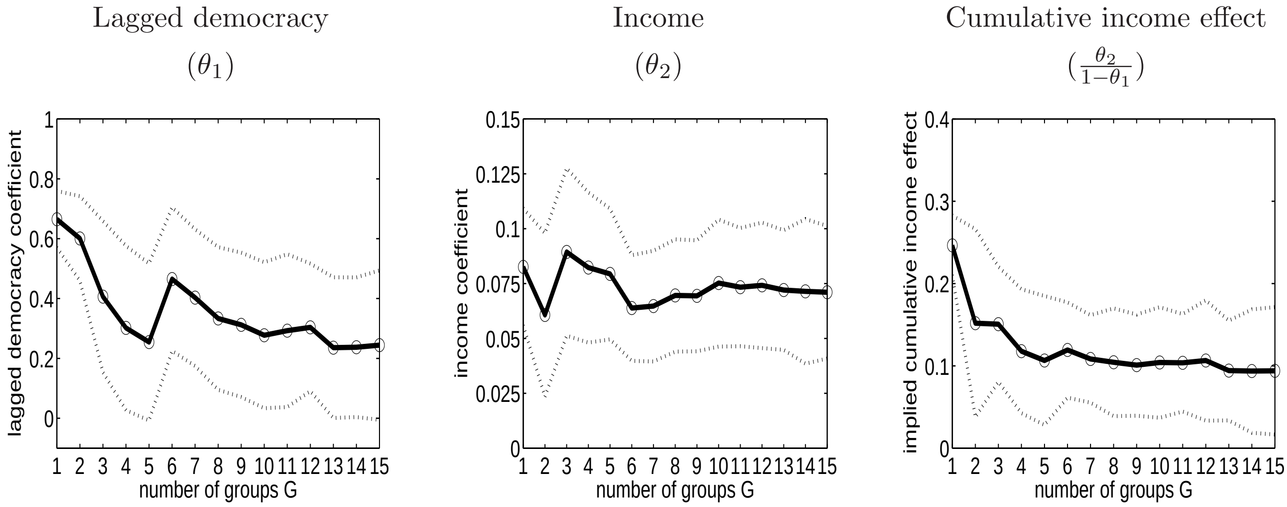
[36] Huntington (1991) distinguishes three waves of democratization: the first one starting in the 1820s in the US and ending with World War I, the second wave lasting between the end of World War II and the early 1960s, and the third wave starting with the Portuguese revolution in 1974. The first two waves were followed by two "counterwaves", in the 1930s and the 1960s, respectively. According to this typology it is still unclear whether the recent Arab spring will be the start of a "fourth wave" of democratization (Diamond, 2011).

[37] All estimates were computed using Algorithm 2– with parameters $(10; 10; 10)$. We performed extensive checks of numerical accuracy, some of which are described in Section 3.

[38] The implied cumulative effect of income shown in Figure 1 is almost identical to the estimated income effect when using a specification that only controls for lagged GDP per capita (and does not include lagged democracy). Results are available upon request.

[39] Table B2 in Appendix B reports three standard error estimates: based on a large-T normal approximation, based on Pollard (1982)'s fixed T normal approximation, and based on the bootstrap (our more conservative estimates, shown in Figure 1). Note that the within-group (that is, within-$(\widehat{g}_i, t)$) variance of income remains sizable as the number of groups increases: it is 65% of the total income variance when $G = 3$, 48% when $G = 10$, and still 43% when $G = 15$. This is substantially larger than the within-country variance of income (6%). In contrast, the within-group variance of

Figure 1: Parameter estimates



Lagged democracy (θ_1)

Income (θ_2)

Cumulative income effect ($\frac{\theta_2}{1-\theta_1}$)

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, computed using Algorithm 2* (5; 10; 5)*.*

is quantitatively small: only 40% of the pooled OLS estimate when $G \geq 5$. Moreover, we will see in Section 6.4 that the association between income and democracy disappears in a specification that combines both time-varying grouped effects and time-invariant country-specific effects.

Lastly, the values reported in Table B2 in Appendix B show that the objective function decreases steadily as G increases: by almost 50% when $G = 5$ compared to OLS, and by 75% when $G = 13$. Interestingly, the last row of the table shows that the objective function of grouped fixed-effects is *lower* than the one of fixed-effects as soon as $G \geq 3$. This suggests that a substantial amount of cross-country heterogeneity is time-varying.[40] We now document these time-varying patterns.

democracy is 10% when $G = 15$, whereas the within-country variance is 26%. This difference arises because the groups are estimated in order to fit the outcome (democracy), but not necessarily the regressor (income).

[40] Another result of Table B2 is that $G = 10$ is optimal according to BIC. Recall from Section 5 that this criterion provides a conservative estimate of the number of groups if T grows at a slower rate than N. Note also that the GFE estimates in Figure 1 do not vary much between $G = 5$ and $G = 15$. According to the discussion in Section 5.2, this is consistent with the true number of groups being actually *smaller* than 10. Optimal choice of G in practice is a notoriously difficult problem in related contexts (e.g., mixture and factor models), which deserves further study.

6.3 Grouped patterns

The 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 $\widehat{\alpha}_{gt}$, as well as average democracy and lagged log-GDP per capita by group over time.

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 will 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 will 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. Note that 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.

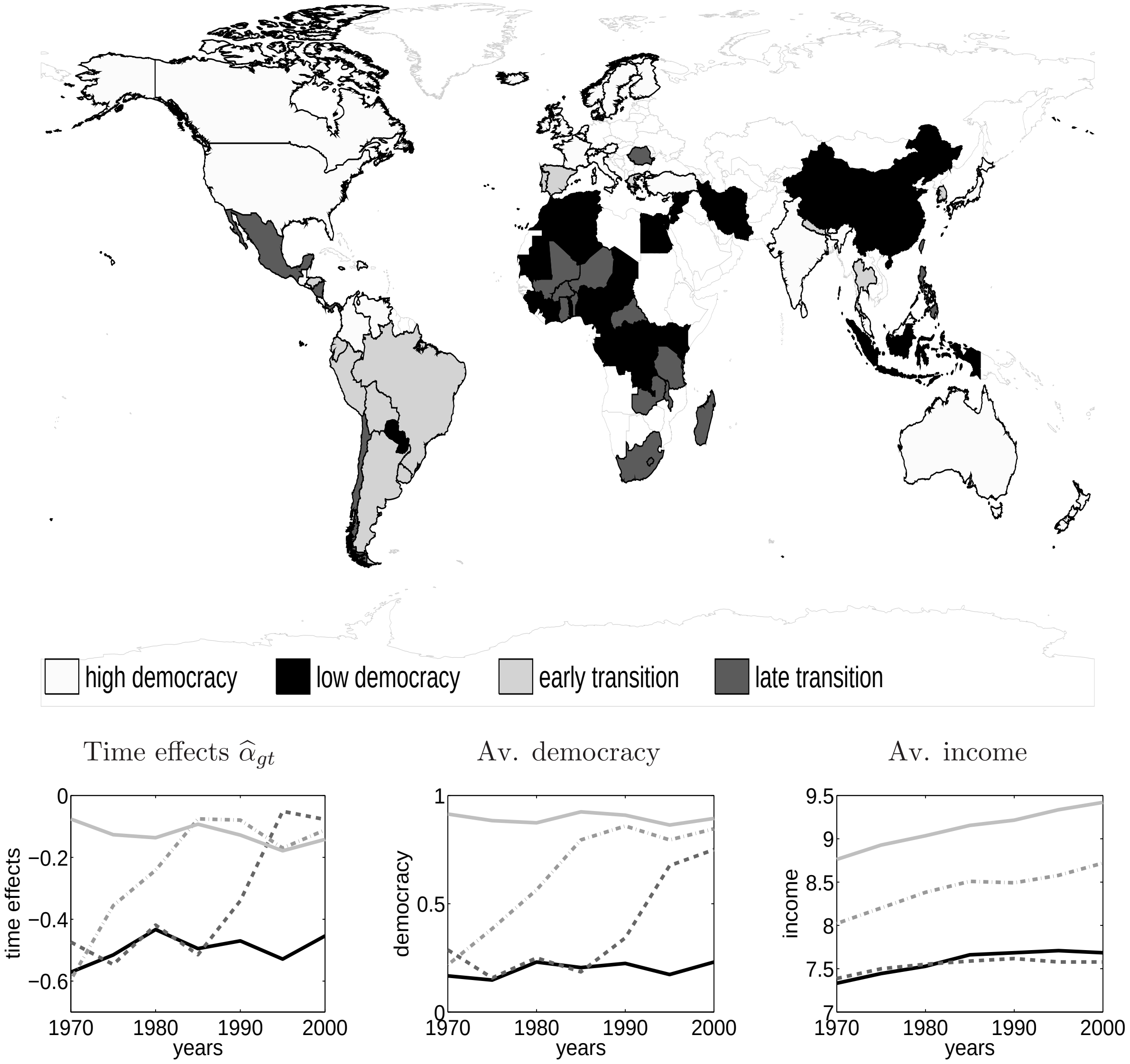
While the first two groups of countries are consistent with FE, the other two are not. Group 3 ("early transition") experiences a marked increase in democracy in the first part of the sample period: its mean Freedom House score increases from .20 in 1970 to almost .90 in 1990. This group includes a large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

In addition to the group-specific means shown in Figure 2, Figure B1 in Appendix B reports uniform 50%-confidence bands for both Freedom House score and lagged log-GDP per capita (thick dashed-dotted lines) for each of the four estimated groups.[41] The figure also shows all country paths of democracy and income over time (thin dotted lines). The left panel shows that, within each group, most countries tend to follow a common group pattern of democracy.[42] At the same time, however, there is evidence of a subtantial amount of heterogeneity in democracy paths, which is only imperfectly captured using the parsimonious 4-groups model. In the next subsection we will present estimates that allow for additional, within-group heterogeneity.

[41] The bands are constructed such that they contain more than 50% of democracy (resp., income) *paths*.

[42] As a complement, Table B4 in Appendix B reports the 1970-2000 evolution of a binary measure of democracy, which classifies as "democratic" (resp., "non-democratic") a country whose Freedom House score is strictly higher (resp., lower) than .50.

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



Note: See the notes to Figure 1. On the bottom panel, the left column reports the group-specific time effects $\widehat{\alpha}_{gt}$. The other two columns 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. Table B6 in Appendix B gives the list of countries by group.

The grouped patterns in Figure 2 remain rather stable as the number of groups changes. Table B6 in Appendix B shows group membership by country, and Figure B2 the corresponding time patterns, for $G = 2, ..., 6$. The specification with $G = 3$ shows two groups essentially identical to Groups 1 and 2 above, and a third one that clusters Groups 3 and 4, which experiences an upward democracy profile over the period. Taking $G = 5$ yields four groups similar to Groups 1-4, plus another group whose democracy level is intermediate between those of Groups 1 and 2, roughly stable over time. This additional group includes Mexico, Indonesia, and Turkey (12 countries in total). When the number of groups is 6 or higher, the estimated group-specific time profiles tend to become more volatile and less easily interpretable.

It is important to note that the time patterns and group membership reported in Figure 2 are estimated from the panel data, and not driven by modelling assumptions other than the group structure. Hence, 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 a result of estimation, not modelling assumptions. In particular, our approach allows group membership and income– our main regressor– to be correlated, in addition to the direct effect of income on democracy which we document in Figure 1.

Although the estimated groups exhibit a strong spatial clustering, they do not match a simple geographic division. To illustrate this, we report in Figure B3 in Appendix B the group-specific time effects and averages of democracy and income, respectively, when the continents are used to form five groups. The results show that, although this simple geographic division yields a clear separation in terms of income and democracy levels, the time patterns are not as clearly separated as in Figure 2. In particular, this specification is not able to distinguish between stable and transition patterns within South America or Africa. In contrast, the grouped fixed-effects estimator selects the grouping that maximizes between-group variation, leading to better identification of stable and transition patterns.

As a different strategy, one could use external data to attempt to classify countries. This is the approach taken by Papaioannou and Siourounis (2008), who combined electoral archives and historical resources for this purpose. Interestingly, their classification of the type of political evolution closely matches the results of GFE estimation.[43] Note that, unlike this data-intensive approach, our automatic method does not require the use of external data.

[43] One of the few clear differences between the classification in Papaioannou and Siourounis (2008) and ours is Iran, which is consistently classified as a "low democracy" country according to our results (e.g., in Group 2), while they classify it as a "borderline" democratization case.

6.4 Robustness checks

We summarize the results of four sets of robustness checks.[44] First, we use the unbalanced panel that covers the period 1960-2000. After dropping all countries with less than 3 observations, we obtain an unbalanced sample of 118 countries.[45] The cumulative income effect is close to the one that we estimated on the balanced sample: for example, it is .13 for $G = 4$ and .12 for $G = 10$. Interestingly, the group classification is very similar between the two samples: when $G = 4$ the group-specific patterns also highlight high and low-democracy countries, as well as early and late transition countries. Moreover, out of the 90 countries of the balanced sample, only 6 change groups when estimated on the unbalanced panel.[46]

As a second check, we follow Acemoglu *et al.* (2008) and use a different measure of democracy: the (normalized) composite Polity index. The balanced panel contains 75 countries, for the same time periods. The grouped fixed-effects estimates are similar to the results obtained using the Freedom House measure. The income effect is .20 in the pooled OLS regression, .06 for GFE with $G = 2$, and decreases slightly to .05 when $G = 15$, significant. Moreover, time patterns and country classification are also similar, although there are some differences related to the measurement of democracy.[47]

As a third check, we include additional controls in model (28). Specifically, following Acemoglu *et al.* (2008) we control for education, log-population size, and age group percentages (5 categories, plus median age). The results are very similar to the main specification. When controlling for education and population size only, the income effect has a similar magnitude ($\approx .10$, significant), while when adding age structure as a control the cumulative income effect drops to .05, marginally significant. For both specifications the time patterns and country classification documented in Figure 2 remain almost unchanged.[48]

As a fourth and important check, we show the results of a model that combines time-varying grouped-specific effects and time-invariant country-specific effects, as in equation (5). The model is

[44] All the results that we refer to in this section, when not directly available in the text or appendix, are available from the authors upon request. Stata codes to replicate the results are available on the first author's webpage: http://www.cemfi.es/∼bonhomme/

[45] The 32 countries we drop using this selection criterion mostly belong to the ex-Republics of the Soviet Union, which became independent in the second part of the sample.

[46] All the countries whose group changes switch from "late" to "early" transition. For example, Mexico, Philippines and Taiwan become part of the early transition countries. As for those countries that are not in the balanced sample: Haiti and Zimbabwe are classified in Group 2 (low democracy), Poland and Hungary in Group 4 (late transition), and Botswana is classified in Group 1 (high-democracy). The world map and group profiles for $G = 4$ corresponding to the unbalanced panel are shown on the first author's webpage.

[47] For example, for $G = 4$ group membership coincides with the one shown in Table B6 in Appendix B except in 11 cases. One of the major disagreements between the two sets of results is South Africa, whose 1980 Polity index is .70, while its Freedom House score is .33.

[48] In both models that control for additional covariates, the BIC criterion selects $G = 7$ groups, a more parsimonious specification than in the case without additional covariates, see footnote 40.

estimated using grouped fixed-effects in deviations to country-specific means. Table B3 in Appendix B shows the estimates of the income effect. According to these results, the implied cumulative effect of income on democracy is insignificant, in contrast with the quantitatively small but statistically significant effect obtained using baseline GFE (see Figure 1). The income effect estimated using FE and GFE at the same time is thus in line with the baseline fixed-effects estimate.

However, the estimated time patterns are remarkably robust to the inclusion of country FE. As we discussed in Section 2, our approach allows to consistently estimate group membership even in the presence of country-specific fixed-effects. The upper panel in Figure B4 in Appendix B shows that a specification allowing for three different types of time patterns in addition to the country-specific fixed-effects yields a similar division between "stable", "early transition", and "late transition" countries. Moreover, the last column in Table B6 shows that the match with the classification without country FE and $G = 4$ is perfect for 80 out of the 90 countries, the "stable" group mostly comprising countries that belonged to Groups 1 and 2 in the baseline specification (see Figure 2). We also estimated the model without including lagged democracy as a control, in order to alleviate potential concerns relative to the presence of the lagged outcome, and found very similar results. Indeed, remarkably similar time profiles and group classifications emerge when using the standard kmeans algorithm (without covariates), in levels and in deviations to country-specific means.[49]

As a last exercise, we experimented with the two-layer model of unobserved heterogeneity (7). This model has G_1 groups with time-varying patterns, and within each of these groups it has G_2 subgroups whose time patterns differ from the common one by an intercept shift. The two-layer model is more parsimonious than the one that combines GFE and FE, and may be well adapted given the short length of the panel. We found it useful to allow for a different number of subgroups within each group, and assume the following two-layer group structure:

$$(g_1, g_2) \in \{(1,1),(1,2),(1,3),(1,4),(1,5),(2,1),(2,2),(3,1),(3,2)\}.$$

The lower panel of Figure B4 in Appendix B shows the time-varying group-specific patterns, and the next-to-last two columns in Table B6 show group membership by country. We see that the two-layer model delivers a clear separation between stable countries, early transition countries, and late transition countries. This output is similar to the baseline GFE specification with $G = 4$, and to the estimates in deviations to country-specific means. Note that the two-layer specification and the latter one deliver almost identical group classifications (except in 5 cases).

In addition, the results provide evidence that the three time-varying groups are heterogeneous themselves. Stable countries show the highest degree of heterogeneity, with 5 subgroups: high

[49] We also estimated the model in first differences. One issue with the first-differenced data is the presence of a mass point at zero for almost 60% of observations in the balanced panel when using the Freedom House measure of democracy. Although the results show some discrepancies with our baseline group classification, particularly for the early transition group, they similarly highlight the presence of three types of time profiles: stable, early and late transition.

democracy countries (such as the US, Japan, Western Europe), medium-high democracy (Colombia, Venezuela), intermediate (Turkey, Malaysia), medium-low (Paraguay, Indonesia, Egypt), and low democracy countries (China, Iran). Early transition countries are divided into high (Spain, Portugal) and low (part of Latin America) democracy levels. Similarly, late transition countries are also divided into high (South Africa, Panama) and low (part of Sub-saharian Africa). Note that the fact that stable countries are separated into 5 subgroups, whereas early and late transition countries are divided into 2 subgroups each, is a result of estimation, not of modelling assumptions.

Overall, the evidence obtained suggests that the income effect 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, at different points in time. In the last part of this section we attempt to find an explanation for why these groups of countries have evolved so differently.

6.5 Explaining the estimated grouped patterns

The country classification of Figure 2 seems to be a robust feature of the democracy/income relationship in the last third of the twentieth century. We now attempt to identify factors that explain why these four estimated groups of countries are associated with such different levels and evolution of democracy and income during this period.

The first set of factors we consider are long-run, historical determinants. Following Acemoglu *et al.* (2008), we use a measure of constraints on the executive at independence, the rationale being that more stringent constraints may be beneficial to embark on a pro-growth, pro-democracy development path. We also consider the date of independence, and a measure of log GDP per capita in 1500, as potential long-run determinants. In addition, we consider the initial democracy level (in 1965), as well as two factors that have been emphasized by the "modernization" theory (Lipset, 1959): log-GDP per capita (in 1965), and a measure of education (average years of schooling, in 1970). We also include shares of Catholic and Protestant in the population (in 1980).

Table B5 in Appendix B shows descriptive statistics by group. Both the high-democracy countries (Group 1) and the early transition ones (Group 3) became independent in the nineteenth century on average, while the countries in the two other groups became independent more recently. The high-democracy group had more stringent constraints on the executive at the time of independence. This group also has a higher initial democracy level in 1965,[50] higher initial income and education, and

[50] Note that the group averages of democracy in 1965 are higher for Groups 2-4 than the 1970 levels that can be seen on Figure 2. This reflects the fact that the 1960s were characterized by a number of transitions to *autocracy*, a feature that we also observed on our estimates from the 1960 − 2000 unbalanced sample.

Table 5: Explaining group membership

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Group 1: high-democracy (vs Group 2: low democracy)									
log GDP p.c. (1500)	1.39 (.971)	.865 (1.74)	.698 (1.76)	–	–	–	−.224 (2.41)	−.307 (2.61)	−.465 (2.75)	−.628 (2.67)
Independence Year/100	–	−4.55 (1.22)	−4.44 (1.27)	–	–	–	−3.51 (1.43)	−3.72 (1.56)	−3.68 (1.75)	−3.59 (1.75)
Constraints	–	7.26 (2.00)	7.12 (2.06)	–	–	–	5.67 (2.49)	4.74 (2.60)	4.70 (2.62)	4.52 (2.77)
Democracy (1965)	–	–	–	7.10 (2.11)	5.80 (2.56)	5.92 (2.66)	–	6.72 (3.39)	6.81 (3.44)	6.24 (3.65)
log GDP p.c. (1965)	–	–	–	1.51 (.587)	–	1.09 (.883)	–	–	.194 (1.25)	.447 (1.35)
Education (1970)	–	–	–	–	.798 (.324)	.492 (.402)	.949 (.373)	.443 (.435)	.418 (.536)	.258 (.560)
Share Catholic (1980)	–	–	.611 (1.20)	–	–	–	–	–	–	−.627 (1.70)
Share Protestant (1980)	–	–	6.81 (4.37)	–	–	–	–	–	–	3.85 (6.32)
	Group 3: early transition (vs Group 2: low democracy)									
log GDP p.c. (1500)	.959 (1.19)	−.894 (1.85)	−.504 (1.87)	–	–	–	−1.19 (2.44)	−2.27 (2.56)	−3.48 (3.03)	−3.13 (2.97)
Independence Year/100	–	−3.53 (1.11)	−3.32 (1.23)	–	–	–	−2.72 (1.23)	−2.96 (1.30)	−4.02 (1.63)	−3.82 (1.76)
Constraints	–	2.25 (2.10)	2.23 (2.34)	–	–	–	.939 (2.47)	.473 (2.56)	.070 (2.57)	.010 (2.95)
Democracy (1965)	–	–	–	−.232 (1.69)	−1.63 (2.03)	−1.79 (2.08)	–	−1.36 (3.03)	−.831 (3.02)	−1.37 (3.16)
log GDP p.c. (1965)	–	–	–	1.40 (.567)	–	.503 (.793)	–	–	−1.87 (1.34)	−1.58 (1.42)
Education (1970)	–	–	–	–	.883 (.311)	.749 (.357)	.570 (.361)	.729 (.425)	1.19 (.565)	1.18 (.601)
Share Catholic (1980)	–	–	1.00 (1.22)	–	–	–	–	–	–	−.215 (1.67)
Share Protestant (1980)	–	–	−.552 (7.87)	–	–	–	–	–	–	−1.55 (8.93)
	Group 4: late transition (vs Group 2: low democracy)									
log GDP p.c. (1500)	−1.06 (1.14)	−.968 (1.07)	−.751 (1.14)	–	–	–	−1.63 (1.95)	−1.97 (2.07)	−1.99 (2.13)	−2.08 (2.16)
Independence Year/100	–	−.681 (.635)	−.785 (.763)	–	–	–	−.027 (.926)	−.144 (.939)	−.219 (1.03)	−.007 (1.38)
Constraints	–	.485 (1.30)	.848 (1.39)	–	–	–	−.607 (1.74)	−1.05 (1.86)	−1.11 (1.88)	−.527 (2.22)
Democracy (1965)	–	–	–	1.23 (1.43)	.047 (1.93)	.134 (1.89)	–	2.39 (2.45)	2.46 (2.45)	1.50 (2.77)
log GDP p.c. (1965)	–	–	–	.021 (.464)	–	−.215 (.701)	–	–	−.263 (.902)	.214 (1.07)
Education (1970)	–	–	–	–	.494 (.302)	.544 (.349)	.597 (.358)	.423 (.389)	.502 (.439)	.331 (.476)
Share Catholic (1980)	–	–	.888 (1.19)	–	–	–	–	–	–	1.20 (1.90)
Share Protestant (1980)	–	–	5.40 (3.87)	–	–	–	–	–	–	5.23 (5.78)

Note: Balanced panel from Acemoglu et al. (2008). "Constraints" are constraints on the executive at independence, measured as in Acemoglu et al. (2005). Multinomial logit regressions of the estimated groups ($G = 4$). The reference group is Group 2 (low-democracy). Group membership is shown on Figure 2. Sample size in the most flexible specification– column (10)– is $N = 68$.

a larger share of Protestant. The early transition group (Group 3) has a higher average education level than the low democracy group, and a larger share of Catholic (63% versus 23%). Lastly, the late transition group (Group 4) differs little from the low democracy one in terms of observables, apart from a slightly higher education level.

In order to jointly assess the effects of the different factors, we next report in Table 5 the results of multinomial logit regressions of the four estimated groups, using several specifications. The large-N, T asymptotic analysis of Section 4 provides a justification for treating the group estimates as data when running the regressions and computing standard errors. The base category is Group 2 (low-democracy). The third row of the top panel of Table 5 shows that constraints on the executive at independence are a significant predictor of the probability of belonging to Group 1 relative to Group 2. This is consistent with the idea that Group 1 and Group 2 countries have embarked on divergent paths at the time of independence, and is suggestive of a very high persistence of early institutions. Note that the effect remains significant at the 10% level even when all other controls (democracy in 1965, income, education...) are included. At the same time, early independence is also associated with a higher likelihood of belonging to Group 1.

However, as shown by the middle and bottom panels of Table 5, constraints on the executive at independence do not significantly affect the probability of belonging to either of the two transition groups (Groups 3 and 4). This suggests that, while conditions at independence partly explain differences between low and high-democracy countries, they do not seem to explain the remarkable evolution of transition countries during the recent period. Education positively affects the probability of belonging to Group 3 relative to Group 2, in line with the "modernization" theory. The date of independence also has a positive effect on the likelihood of belonging to the early transition group.[51] In contrast, the bottom panel of the table shows that none of the determinants that we consider (e.g., education or religion) is able to distinguish late transition countries (Group 4) from low democracy countries (Group 2). Overall, these results point to the need to further study the short and long-run determinants of political development.

7 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 or predetermined covariates, and can allow for unit-specific fixed-effects in addition to the time-varying grouped patterns. Importantly, the relationship between group membership and observed

[51] These results are consistent with Papaioannou and Siourounis (2008), who modelled the probability of democratization of countries that started the period as autocracies. They found little evidence for an effect of early institutions. In addition, their results suggest that more educated societies are more likely to become democratic.

covariates is left unrestricted.

The GFE approach should be useful in applications where time-invariance of the fixed-effects is an implausible assumption, and 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 social interactions models where reference groups are estimated from the data, and models of spatial dependence with an endogenous spatial weights matrix. Computation of the estimator is challenging, but recent advances in the literature on data clustering have allowed us to build an efficient computation routine.

Our asymptotic results show that, though subject to an incidental parameter problem, GFE has attractive large-N, T properties. In particular, there is no need to perform bias reduction. In future work, we plan to study two issues. First, we think it would be useful to better characterize the statistical properties of the GFE estimator when group separation fails. A second interesting extension is the study of the GFE approach in nonlinear models. We are particularly interested in dynamic discrete choice models, where a discrete modelling of unobserved heterogeneity may be appealing (Kasahara and Shimotsu, 2009, Browning and Carro, 2011).

The next step in our research is to relax the assumption that there is a finite number of groups in the population. As an alternative approach, one may view the grouped model as an approximation to the underlying data generating process, and thus let the number of groups grow with the two dimensions of the panel. We are currently studying the properties of the GFE estimator in this large-N, T, G asymptotic framework.

Lastly, in the context of the empirical study of democracy, our results bring new evidence while leaving many questions unanswered. Our estimates of the association between income and democracy are consistent with the results of Acemoglu *et al.* (2008), and suggest that the income effect is plausibly very small or zero in this dataset. However, at the same time the GFE results highlight the presence of "wave" patterns of democratization during the last part of the twentieth century. This raises interesting questions for the political economy literature. Which factors explain democratic transitions? Why did a large share of low democracy countries– including a substantial share of Africa– make a transition in the 1990s?[52] Finally, and importantly, why do we observe groups of countries making transitions at similar points in time?

[52] Note that most of the late transition countries in Figure 2 are Sub-saharian African countries, which made democratic transitions in the 1990s. Interestingly, Brückner and Ciccone (2011) document an association between drought and posterior increases in democracy levels in Sub-saharian Africa. They interpret this evidence as suggesting that a fall in *transitory* income may foster democratic change.

References

[1] Acemoglu, D., S. Johnson, and J. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91(5), 1369–1401.

[2] Acemoglu, D., S. Johnson, and J. Robinson (2005): "The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth," *American Economic Review*, 95, 546–79.

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

[4] Aloise, D., P. Hansen, L. and Liberti (2010): "An Improved Column Generation Algorithm for Minimum Sum-of- Squares Clustering," *Mathematical Programming, Ser. A*, DOI 10.1007/s10107-010-0349-7.

[5] 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.*

[6] Alvarez, J. and M. Arellano (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators", *Econometrica*, 71, 1121–1159.

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

[8] Arellano, M. (1987): "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49(4), 431–434.

[9] Arellano, M., and S. Bonhomme (2009): "Robust Priors in Nonlinear Panel Data Models", *Econometrica*, 77, 489–536.

[10] Arellano, M., and S. Bonhomme (2011): "Identifying Distributional Characteristics in Random Coefficients Panel Data Models," to appear in the *Review of Economic Studies.*

[11] 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.

[12] Bai, J. (1994): "Least Squares Estimation of Shift in Linear Processes," *Journal of Time Series Analysis*, 15, 453–472.

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

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

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

[16] Bai, J., and S. Ng (2006): “Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions,” *Econometrica*, 74, 1133–1150.

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

[18] Bester, A., and C. Hansen (2010): “Grouped Effects Estimators in Fixed Effects Models”, unpublished manuscript.

[19] 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.

[20] Browning, M., and J. Carro (2007): “Heterogeneity and Microeconometrics Modelling,” in Advances in Economics and Econometrics, Theory and Applications: Ninth World Congress of the Econometric Society, Vol. 3, ed. by R. Blundell, W. Newey, and T. Persson. Cambridge, U.K.: Cambridge University Press, 47–74.

[21] Browning, M., and J. Carro (2011): “Dynamic Binary Outcome Models with Maximal Heterogeneity”, unpublished manuscript.

[22] Brückner, M. and A. Ciccone (2011): “Rain and the Democratic Window of Opportunity,” *Econometrica*, 79(3), 923–947.

[23] Brusco, M.J. (2006): “A Repetitive Branch-and-Bound Procedure for Minimum Within-Cluster Sums of Squares Partitioning,” *Psychometrika*, 71, 357–373.

[24] Brusco, M.J., and D. Steinley (2007): “A Comparison of Heuristic Procedures for Minimum Within-Cluster Sums of Squares Partitioning,” *Psychometrika*, 72(4), 583–600.

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

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

[27] Caporossi, G., and P. Hansen (2005): “Variable Neighborhood Search for Least Squares Clusterwise Regression,” Cahiers du Gerad G200561.

[28] Diamond, L. (2011): “A Fourth Wave or False Start? Democracy After the Arab Spring,” *Foreign Affairs*, May 22.

[29] Du Merle, O., P. Hansen, B. Jaumard, and N. Mladenovic (2001): "An Interior Point Method for Minimum Sum-of-Squares Clustering," *SIAM J. on Scientific Computing*, 21, 1485–1505.

[30] Durlauf, S.N., Kourtellos, A. and Minkin, A. (2001): "The Local Solow Growth Model," *European Economic Review*, 45(46), 928–940.

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

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

[33] Geweke, J. and M. Keane (2007): "Smoothly Mixing Regressions," *Journal of Econometrics*, 138(1), 252–290.

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

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

[36] Hahn, J., and W. Newey (2004): "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models," *Econometrica*, 72, 1295–1319.

[37] Hansen, C. (2007): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data when T is Large," *Journal of Econometrics*, 141(2), 597–620.

[38] 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.

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

[40] Heckman, J. (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109, 673–748.

[41] Heckman, J., and B. Singer (1984): "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 52(2), 271–320.

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

[43] Inaba, M., N. Katoh, and H. Imai (1994): "Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering," in *Proceedings of the 10th Annual Symposium on Computational Geometry.* ACM Press, Stony Brook, USA, 332–339

[44] Kane, T.J., D.O. Staiger (2002): "The Promise and Pitfalls of Using Imprecise School Accountability Measures," *Journal of Economic Perspectives*, 16(4), 91–114.

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

[46] Keane, M.P., and K.I. Wolpin (1997): "The Career Decisions of Young Men," *Journal of Political Economy*, 105(3), 473–522.

[47] Lin, C. C., and S. Ng (2011): "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown", to appear in *Journal of Econometric Methods*.

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

[49] Maitra, R., A. D. Peterson, and A. P. Ghosh (2011): "A Systematic Evaluation of Different Methods for Initializing the Clustering Algorithm," unpublished working paper.

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

[51] 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.

[52] Moon, H., and M. Weidner (2010a): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," unpublished manuscript.

[53] Moon, H., and M. Weidner (2010b): "Linear Regression for Panel with Unknown Number of Factors as Interactive Fixed Effects," unpublished manuscript.

[54] Munshi, K., and M. Rosenzweig (2009): "Why is Mobility in India so Low? Social Insurance, Inequality, and Growth," BREAD Working Paper No. 092.

[55] Newey, W.K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *in* R.F. Engle and D.L. McFadden, eds., *Handbook of Econometrics* vol 4: 2111-245. Amsterdam: Elsevier Science.

[56] Newey, W.K., and K. D. West (1987): "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

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

[58] Norets, A. (2010): “Approximation of Conditional Densities by Smooth Mixtures of Regressions,” *Annals of Statistics*, 38(3), 1733–1766.

[59] Pacheco, J. and O. Valencia (2003): “Design of Hybrids for the Minimum Sum-of-Squares Clustering Problem,” *Computational Statistics and Data Analysis*, 43(2), 235–248.

[60] Papaioannou, E., and G. Siourounis (2008): “Economic and Social Factors Driving the Third Wave of Democratization,” *Journal of Comparative Economics*, 36, 365–387.

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

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

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

[64] Przeworski, A., M. Alvarez, J. A. Cheibub, and F. Limongi (2000): *Democracy and Development: Political Institutions and Material Well-being in the World, 19501990*. New York: Cambridge University Press.

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

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

[67] Schulhofer-Wohl, S. (2011): “Heterogeneity and Tests of Risk Sharing,” *Journal of Political Economy*, 119, 925–58.

[68] Späth, H. (1979): “Algorithm 39: Clusterwise linear regression,” *Computing*, 22(4), 367–373.

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

[70] 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.

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

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

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

APPENDIX

A Proofs

A.1 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:

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

We also define the following auxiliary objective function:

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

We start by showing the following uniform convergence result.

Lemma A1 *Let Assumption 1.a-1.g 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.

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

We now show that the three terms on the right-hand side of this equality are $o_p(1)$, uniformly on the parameter space.

- By Assumption 1.e 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 $\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}x_{it} = o_p(1)$, uniformly on the parameter space. In addition, $\left\|\theta^0 - \theta\right\|$ is bounded by Assumption 1.a.

- By the CS inequality:

$$\begin{aligned}\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T v_{it}\alpha^0_{g_i^0t}\right)^2 &\leq \frac{1}{N}\sum_{i=1}^N\left(\frac{1}{T}\sum_{t=1}^T v_{it}\alpha^0_{g_i^0t}\right)^2\\ &= \sum_{g=1}^G\frac{1}{N}\sum_{i=1}^N \mathbf{1}\left\{g_i^0=g\right\}\left(\frac{1}{T}\sum_{t=1}^T v_{it}\alpha^0_{gt}\right)^2\\ &\leq \sum_{g=1}^G\frac{1}{N}\sum_{i=1}^N\left(\frac{1}{T}\sum_{t=1}^T v_{it}\alpha^0_{gt}\right)^2\\ &= \sum_{g=1}^G\frac{1}{NT^2}\sum_{i=1}^N\sum_{t=1}^T\sum_{s=1}^T v_{it}v_{is}\alpha^0_{gt}\alpha^0_{gs},\end{aligned}$$

which by Assumption 1.d is bounded in expectation by a constant divided by T. This implies that $\frac{1}{NT}\sum_{i=1}^N\sum_{t=1}^T v_{it}\alpha^0_{g_i^0t}$ is uniformly $o_p(1)$.

- Finally we have:

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

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

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

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

Now, note that:

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

By Assumption 1.f:

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

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.g.

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

■

We will also need the following result, which shows that $\widetilde{\mathcal{Q}}(\cdot)$ is maximized at true values.

Lemma A2 *We have, 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.h.

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, from standard least-squares algebra:

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

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

■

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

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

So, by Lemma A2 we have:

$$\widehat{\rho}\left\|\widehat{\theta} - \theta^0\right\|^2 = o_p(1),$$

so it follows from Assumption 1.h that $\left\|\widehat{\theta} - \theta^0\right\|^2 = o_p(1)$.

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

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

which is $o_p(1)$ by Assumptions 1.a and 1.b, and by consistency of $\widehat{\theta}$.

Combining with (A2) we obtain:

$$\widetilde{Q}\left(\theta^0,\widehat{\alpha},\widehat{\gamma}\right) \quad \leq \quad \widetilde{Q}\left(\theta^0,\alpha^0,\gamma^0\right)+o_p(1),$$

from which it follows that:

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

This completes the proof of Theorem 1.

A.2 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 in $\mathbb{R}^{GT}$:

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

We have the following result.[53]

Lemma A3 *Let Assumptions 1.a-1.h, 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.

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

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

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

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

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

Now:

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

[53] Note that group separation (Assumption 2.b) is assumed to show Lemma A3. Proving consistency of the group-specific time effects absent this assumption would require different arguments.

which tends to zero in probability by Theorem 1.

We have thus shown that, for all g:

$$\min_{\widetilde{g}\in\{1,...,G\}} \frac{1}{T}\sum_{t=1}^{T}\left(\widehat{\alpha}_{\widetilde{g}t}-\alpha^0_{gt}\right)^2 \xrightarrow{p} 0. \tag{A3}$$

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

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

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 triangular inequality we have:

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

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 (A3). 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 $g \in \{1, ..., G\}$:

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

where we have used (A3), and the fact that $\widetilde{g} = \sigma\left[\sigma^{-1}\left(\widetilde{g}\right)\right]$.

This completes the proof.

■

The proof of Lemma A3 shows that there exists a permutation $\sigma : \{1, ..., G\} \to \{1, ..., G\}$ such that:

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

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 $\eta > 0$, 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 $\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}-\alpha^0_{gt}\right)^2 < \eta$ for all $g \in \{1, ..., G\}$. We have the following result, which shows that the probability that $\widehat{g}_i(\theta, \alpha)$ differs from g_i^0 tends to zero at a faster-than-polynomial rate, provided (θ, α) is taken in a small enough neighborhood $\mathcal{N}_\eta$.

Lemma A4 *For $\eta > 0$ small enough we have, for all $\delta > 0$ and as T tends to infinity:*

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

Proof.

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

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

So:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)\neq g_i^0\right\} &= \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left\{g_i^0\neq g\right\}\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\\ &\leq \sum_{g=1}^{G}\frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathbf{1}\{g_i^0\neq g\}\mathbf{1}\left\{\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{gt}\right)^2 \leq \sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{g_i^0t}\right)^2\right\}}_{Z_{ig}(\theta,\alpha)}.\end{aligned} \tag{A4}$$

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:

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

Let us now define:

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

We have:

$$\begin{aligned}A_T \leq\ & \underbrace{\left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)v_{it}-\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}\right|}_{=A_{1T}}+\underbrace{\left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)x_{it}'\left(\theta^0-\theta\right)\right|}_{=A_{2T}}\\ &+\underbrace{\left|\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}-\alpha_{gt}\right)\left(\alpha^0_{\widetilde{g}t}-\frac{\alpha_{gt}+\alpha_{\widetilde{g}t}}{2}\right)-\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)\left(\alpha^0_{\widetilde{g}t}-\frac{\alpha^0_{gt}+\alpha^0_{\widetilde{g}t}}{2}\right)\right|}_{=A_{3T}}.\end{aligned}$$

We now bound each of the three terms, for $(\theta,\alpha)\in\mathcal{N}_\eta$.

- We have, by the Cauchy-Schwartz inequality:

$$\begin{aligned} A_{1T} &= \left|\sum_{t=1}^{T}\left[(\alpha_{\widetilde{g}t}-\alpha_{gt})-\left(\alpha_{\widetilde{g}t}^{0}-\alpha_{gt}^{0}\right)\right]v_{it}\right| \\ &\leq T\left(\frac{1}{T}\sum_{t=1}^{T}\left[(\alpha_{\widetilde{g}t}-\alpha_{gt})-\left(\alpha_{\widetilde{g}t}^{0}-\alpha_{gt}^{0}\right)\right]^{2}\right)^{\frac{1}{2}}\times\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^{2}\right)^{\frac{1}{2}} \\ &\leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^{2}\right)^{\frac{1}{2}}, \end{aligned}$$

where C_1 is independent of η and T, and where we have used the definition of $\mathcal{N}_\eta$.

- Next we have, by the CS inequality:

$$\begin{aligned} A_{2T} &= \left|\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}-\alpha_{gt})\,x_{it}'\left(\theta^0-\theta\right)\right| \\ &\leq \sum_{t=1}^{T}\left|(\alpha_{\widetilde{g}t}-\alpha_{gt})\,x_{it}'\left(\theta^0-\theta\right)\right| \\ &\leq T\left(2\sup_{\alpha_t\in\mathcal{A}}|\alpha_t|\right)\times\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)\times\left\|\theta^0-\theta\right\| \\ &\leq TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right), \end{aligned}$$

where C_2 is independent of η and T, and where we have used Assumption 1.a.

- Finally we have, by simple rearrangement:

$$\begin{aligned} A_{3T} &= \left|\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}-\alpha_{gt})\left(\alpha_{\widetilde{g}t}^{0}-\frac{\alpha_{gt}+\alpha_{\widetilde{g}t}}{2}\right)-\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^{0}-\alpha_{gt}^{0}\right)\left(\alpha_{\widetilde{g}t}^{0}-\frac{\alpha_{gt}^{0}+\alpha_{\widetilde{g}t}^{0}}{2}\right)\right| \\ &= \left|\sum_{t=1}^{T}\alpha_{\widetilde{g}t}^{0}\left(\alpha_{\widetilde{g}t}-\alpha_{\widetilde{g}t}^{0}-\alpha_{gt}+\alpha_{gt}^{0}\right)+\frac{1}{2}\sum_{t=1}^{T}\left(\left[\alpha_{\widetilde{g}t}^{0}\right]^{2}-\left[\alpha_{\widetilde{g}t}\right]^{2}-\left[\alpha_{gt}^{0}\right]^{2}+\left[\alpha_{gt}\right]^{2}\right)\right|. \end{aligned}$$

It thus follows from the CS inequality and Assumption 1.a that, for $(\theta,\alpha)\in\mathcal{N}_\eta$:

$$A_{3T} \leq TC_3\sqrt{\eta},$$

where C_3 is independent of η and T.

Combining, we obtain that:

$$\begin{aligned} Z_{ig}(\theta,\alpha) &\leq \max_{\widetilde{g}\neq g}\mathbf{1}\left\{\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}-\alpha_{gt})\left(v_{it}+x_{it}'\left(\theta^0-\theta\right)+\alpha_{\widetilde{g}t}^{0}-\frac{\alpha_{gt}+\alpha_{\widetilde{g}t}}{2}\right)\leq 0\right\} \\ &\leq \max_{\widetilde{g}\neq g}\mathbf{1}\left\{\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^{0}-\alpha_{gt}^{0}\right)\left(v_{it}+\alpha_{\widetilde{g}t}^{0}-\frac{\alpha_{gt}^{0}+\alpha_{\widetilde{g}t}^{0}}{2}\right)\leq A_T\right\} \\ &\leq \max_{\widetilde{g}\neq g}\mathbf{1}\Bigg\{\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^{0}-\alpha_{gt}^{0}\right)\left(v_{it}+\alpha_{\widetilde{g}t}^{0}-\frac{\alpha_{gt}^{0}+\alpha_{\widetilde{g}t}^{0}}{2}\right) \\ &\qquad\qquad \leq TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^{2}\right)^{\frac{1}{2}}+TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right)+TC_3\sqrt{\eta}\Bigg\}. \end{aligned}$$

Using that:

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

we obtain:

$$Z_{ig}(\theta,\alpha) \quad \leq \quad \widetilde{Z}_{ig},$$

where:

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

Note that $\widetilde{Z}_{ig}$ does not depend on (θ,α). In particular we also have:

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} Z_{ig}(\theta,\alpha) \leq \widetilde{Z}_{ig},$$

and thus:

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i\left(\theta,\alpha\right)\neq g_i^0\} \leq \frac{1}{N}\sum_{i=1}^{N}\sum_{g=1}^{G}\widetilde{Z}_{ig}. \tag{A5}$$

Fix $\widetilde{M} > \sqrt{M}$, where M is given by Assumption 1. Note that $\mathbb{E}(v_{it}^2) \leq \sqrt{M}$, and $\mathbb{E}(\|x_{it}\|) \leq \sqrt{M}$. We have, using standard probability algebra and for all g:

$$\begin{aligned}\Pr\left(\widetilde{Z}_{ig}=1\right) \quad \leq \quad & \sum_{\widetilde{g}\neq g}\Pr\Bigg(\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it} \leq -\frac{1}{2}\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)^2 + TC_1\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2\right)^{\frac{1}{2}} \\ & \qquad + TC_2\sqrt{\eta}\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|\right) + TC_3\sqrt{\eta}\Bigg) \\ \leq \quad & \sum_{\widetilde{g}\neq g}\Bigg[\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)^2 \leq \frac{c_{g,\widetilde{g}}}{2}\right) + \Pr\left(\frac{1}{T}\sum_{t=1}^{T}v_{it}^2 \geq \widetilde{M}\right) + \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\| \geq \widetilde{M}\right) \\ & + \Pr\left(\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it} \leq -T\frac{c_{g,\widetilde{g}}}{4} + TC_1\sqrt{\eta}\sqrt{\widetilde{M}} + TC_2\sqrt{\eta}\widetilde{M} + TC_3\sqrt{\eta}\right)\Bigg].\end{aligned} \tag{A6}$$

To end the proof of Lemma A4, 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).

Theorem A1 *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 $s^2 = \sup_{t\geq 1}\left(\sum_{s\geq 1}|\mathbb{E}\left(z_t z_s\right)|\right) < \infty$. Let also $d = \frac{d_1 d_2}{d_1+d_2}$. By evaluating inequality (1.7) in Merlevède, Peligrad and Rio (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 $T \geq 1$:

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

Theorem A1 directly follows.

■

To deal with the case where lagged outcomes $y_{i,t-1}$ are included as additional covariates, we will use the following corollary of Theorem A1. We provide a separate result in this case, as there exist simple examples of autoregressive processes that fail to be strongly mixing (e.g., Andrews, 1984). For simplicity we only deal with first-order autoregressive structures.

Corollary A1 *Let z_t be a process that satisfies the tail and dependence conditions of Theorem A1, and let:*

$$\omega_t = \rho\omega_{t-1} + z_t, \quad t = 1, ..., T,$$

where $|\rho| < 1$, and where ω_0 satisfies the tail condition of Theorem A1. Let us suppose that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(|z_t|\right) \leq C$ is bounded. Then, for all $\omega > \frac{2C}{1-|\rho|}$ we have, for all $\delta > 0$:

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

Proof. We have: $\omega_t = \rho^t\omega_0 + \sum_{s=0}^{t-1}\rho^s z_{t-s}$. Hence:

$$\begin{aligned}\Pr\left(\frac{1}{T}\sum_{t=1}^{T}|\omega_t| \geq \omega\right) &\leq \Pr\left(\frac{1}{T}\sum_{t=1}^{T}|\rho|^t|\omega_0| \geq \frac{\omega}{2}\right) + \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{s=0}^{t-1}|\rho|^s|z_{t-s}| \geq \frac{\omega}{2}\right)\\ &\leq \Pr\left(|\omega_0| \geq T\frac{\omega}{2}(1-|\rho|)\right) + \Pr\left(\frac{1}{T}\sum_{m=1}^{T}|z_m| \geq \frac{\omega}{2}(1-|\rho|)\right),\end{aligned}$$

where we have used the change-in-variables $m = t - s$. The first term on the right-hand side is $o\left(T^{-\delta}\right)$ by the tail condition, whereas the second term is $o\left(T^{-\delta}\right)$ by Theorem A1 applied to $|z_t| - \mathbb{E}(|z_t|)$ and taking $z = \frac{\omega}{2}(1-|\rho|) - C > 0$.

This ends the proof of Corollary A1.

■

We now bound the four terms on the right-hand side of (A6).

- 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 Theorem A1 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 second term on the right-hand side of (A6), applying Theorem A1 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.

• As for the third term there are two cases, depending on whether part (i) or (ii) is satisfied in Assumption 2.e. If part (i) holds then by choosing $\widetilde{M}$ larger than the upper bound on $\|x_{it}\|$ yields $\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\| < \widetilde{M}$.

If instead part (ii) in Assumption 2.e holds, then we apply Theorem A1 to $z_t = \|x_{it}\| - \mathbb{E}(\|x_{it}\|)$ and take $z = \widetilde{M} - \sqrt{M}$, yielding:

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

Moreover, if a lagged outcome $y_{i,t-1}$ is included as an additional regressor, then one can apply Corollary A1 with $\omega_t = y_{i,t-1}$. The assumptions of the Corollary are satisfied provided the initial condition y_{i0} has faster-than-polynomial tails. Let $C \geq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left(|x_{it}'\theta^0 + \alpha_{g_i^0 t}^0 + v_{it}|\right)$ be a uniform constant, and let us take $\widetilde{M} > 2C/\left(1-|\rho^0|\right)$, where ρ^0 is the autoregressive coefficient. One then has: $\Pr\left(\frac{1}{T}\sum_{t=1}^{T}|y_{i,t-1}| \geq \widetilde{M}\right) = o\left(T^{-\delta}\right)$.

• Lastly, to bound the fourth term on the right-hand side of (A6) we denote as c the minimum of $c_{g,\widetilde{g}}$ over all $g \neq \widetilde{g}$ and we take:

$$\eta \leq \left(\frac{c}{8\left(C_1\sqrt{\widetilde{M}} + C_2\widetilde{M} + C_3\right)}\right)^2. \tag{A7}$$

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

Taking η satisfying (A7) yields, for all $\widetilde{g} \neq g$:

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

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

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

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

Lastly, applying Theorem A1 with $z_t = \left(\alpha_{\widetilde{g}t}^0 - \alpha_{gt}^0\right)v_{it}$ and taking $z = \frac{c_{g,\widetilde{g}}}{8}$ yields:

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

Combining the results we finally obtain, using (A6), that for η satisfying (A7) and for all $\delta > 0$:

$$\Pr\left(\widetilde{Z}_{ig} = 1\right) \quad = \quad o\left(T^{-\delta}\right).$$

Moreover, noting that the above upper bounds on the probabilities do not depend on i and g we have:

$$\sup_{i\in\{1,\ldots,N\},g\in\{1,\ldots,G\}} \Pr\left(\widetilde{Z}_{ig} = 1\right) \quad = \quad o\left(T^{-\delta}\right). \tag{A8}$$

To complete the proof of Lemma A4 note that, for η that satisfies (A7) we have, for all $\delta > 0$ and all $\varepsilon > 0$:

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

where we have used (A5), the Markov inequality, and (A8), respectively.

This ends the proof of Lemma A4.

■

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:[54]

$$\widehat{Q}(\theta,\alpha)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2, \tag{A9}$$

and:

$$\widetilde{Q}(\theta,\alpha)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{g_i^0t}\right)^2. \tag{A10}$$

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

By the CS inequality we have:

$$\begin{aligned}
\left(\widehat{Q}(\theta,\alpha)-\widetilde{Q}(\theta,\alpha)\right)^2 &\leq \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0t}-\alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2\times\dots\\
&\qquad \frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\frac{\alpha_{\widehat{g}_i(\theta,\alpha)t}+\alpha_{g_i^0t}}{2}\right)^2,
\end{aligned}$$

where the second term on the right-hand side is uniformly $O_p(1)$ by Assumptions 1.a-1.c.

Now we have:

$$\begin{aligned}
\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\alpha_{g_i^0t}-\alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2 &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\}\left(\alpha_{g_i^0t}-\alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2\\
&\leq \left(4\sup_{\alpha_t\in\mathcal{A}}\alpha_t^2\right)\times\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i(\theta,\alpha)\neq g_i^0\}.
\end{aligned}$$

Let $\eta > 0$ be small enough such that Lemma A4 is satisfied. Using the two above inequalities, Assumption 1.a, and Lemma A4 we have, for all $\delta > 0$:

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\left|\widehat{Q}(\theta,\alpha)-\widetilde{Q}(\theta,\alpha)\right|=o_p\left(T^{-\delta}\right). \tag{A11}$$

[54] 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.

Now, by consistency of $\widehat{\theta}$ (Theorem 1) and $\widehat{\alpha}$ (Lemma A3) we have, as N and T tend to infinity:

$$\Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right)\notin\mathcal{N}_\eta\right)\to 0. \tag{A12}$$

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

$$\Pr\left(\left(\widetilde{\theta},\widetilde{\alpha}\right)\notin\mathcal{N}_\eta\right)\to 0. \tag{A13}$$

Combining (A11) and (A12) we have, for all $\delta>0$ and as N and T tend to infinity:

$$\widehat{Q}\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)=o_p\left(T^{-\delta}\right). \tag{A14}$$

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

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

which is $o(1)$ by (A11) and (A12).

Similarly, combining (A11) and (A13) we obtain:

$$\widehat{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right)-\widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) \quad=\quad o_p\left(T^{-\delta}\right). \tag{A15}$$

Next, note that, by the definition of $\left(\widetilde{\theta},\widetilde{\alpha}\right)$:

$$\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right)\geq 0.$$

Moreover, using (A14), (A15), and the definition of $\left(\widehat{\theta},\widehat{\alpha}\right)$ yields:

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

It thus follows that:

$$\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) \quad=\quad o_p\left(T^{-\delta}\right). \tag{A16}$$

Now, we have:

$$\begin{aligned}\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) &= \frac{2}{NT}\sum_{i=1}^N\sum_{t=1}^T\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\widetilde{\alpha}_{g_i^0t}-\widehat{\alpha}_{g_i^0t}\right)\left(y_{it}-x_{it}'\left(\frac{\widetilde{\theta}+\widehat{\theta}}{2}\right)-\frac{\widetilde{\alpha}_{g_i^0t}+\widehat{\alpha}_{g_i^0t}}{2}\right)\\ &= \left(\widetilde{\theta}-\widehat{\theta}\right)'\frac{2}{NT}\sum_{i=1}^N\sum_{t=1}^T x_{it}\left(y_{it}-x_{it}'\left(\frac{\widetilde{\theta}+\widehat{\theta}}{2}\right)-\frac{\widetilde{\alpha}_{g_i^0t}+\widehat{\alpha}_{g_i^0t}}{2}\right)\\ &\quad+\frac{1}{T}\sum_{g=1}^G\sum_{t=1}^T\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\frac{2}{N}\sum_{i=1}^N\mathbf{1}\{g_i^0=g\}\left(y_{it}-x_{it}'\left(\frac{\widetilde{\theta}+\widehat{\theta}}{2}\right)-\frac{\widetilde{\alpha}_{gt}+\widehat{\alpha}_{gt}}{2}\right).\end{aligned} \tag{A17}$$

Note that, as $\left(\widetilde{\theta},\widetilde{\alpha}\right)$ is a least squares estimator, the following empirical moment restrictions are satisfied:

$$\begin{aligned}\frac{2}{NT}\sum_{i=1}^N\sum_{t=1}^T x_{it}\left(y_{it}-x_{it}'\widetilde{\theta}-\widetilde{\alpha}_{g_i^0t}\right) &= 0\\ \frac{2}{N}\sum_{i=1}^N\mathbf{1}\{g_i^0=g\}\left(y_{it}-x_{it}'\widetilde{\theta}-\widetilde{\alpha}_{gt}\right) &= 0, \quad \text{for all } (g,t).\end{aligned}$$

Combining with (A17) yields:

$$\begin{aligned}\widetilde{Q}\left(\widehat{\theta},\widehat{\alpha}\right)-\widetilde{Q}\left(\widetilde{\theta},\widetilde{\alpha}\right) &= \left(\widetilde{\theta}-\widehat{\theta}\right)'\frac{2}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}x_{it}\left(x_{it}'\left(\frac{\widetilde{\theta}-\widehat{\theta}}{2}\right)+\sum_{g=1}^{G}\mathbf{1}\{g_i^0=g\}\left(\frac{\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}}{2}\right)\right)\\ &\quad +\frac{1}{T}\sum_{g=1}^{G}\sum_{t=1}^{T}\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\frac{2}{N}\sum_{i=1}^{N}\mathbf{1}\{g_i^0=g\}\left(x_{it}'\left(\frac{\widetilde{\theta}-\widehat{\theta}}{2}\right)+\frac{\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}}{2}\right) \qquad \text{(A18)}\\ &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\widetilde{\theta}-\widehat{\theta}\right)+\sum_{g=1}^{G}\mathbf{1}\{g_i^0=g\}\left(\widetilde{\alpha}_{gt}-\widehat{\alpha}_{gt}\right)\right)^2,\end{aligned}$$

so that:

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

It thus follows that:

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

where $\widehat{\rho}\xrightarrow{p}\rho>0$ as a consequence of Assumption 1.h.

Hence, $\widetilde{\theta}-\widehat{\theta}=o_p\left(T^{-\delta}\right)$ for all $\delta>0$. This shows (17).

Properties of $\widehat{\alpha}$. Using (A18) above, consistency of $\widehat{\theta}$ and $\widetilde{\theta}$, Assumptions 1.a and 1.b, and equation (A16), we obtain:

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

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

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

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:[55]

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

Now we have, taking η such that (A7) is satisfied:

$$\Pr\left(\left(\widehat{\theta},\widehat{\alpha}\right)\notin\mathcal{N}_\eta\right) = o(1).$$

We also have, with probability one:

$$\begin{aligned}\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\Pr\left(\sup_{i\in\{1,...,N\}}\left|\widehat{g}_i\left(\theta,\alpha\right)-g_i^0\right|>0\right) &\leq N\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\sup_{i\in\{1,...,N\}}\Pr\left(\left|\widehat{g}_i\left(\theta,\alpha\right)-g_i^0\right|>0\right)\\ &= N\sup_{i\in\{1,...,N\}}\sup_{(\theta,\alpha)\in\mathcal{N}_\eta}\Pr\left(\widehat{g}_i\left(\theta,\alpha\right)\neq g_i^0\right).\end{aligned}$$

[55] Note that the neighborhood $\mathcal{N}_\eta$ depends on the processes $\{\alpha_{gt}^0\}_t$, for $g=1,...,G$.

Moreover, the proof of Lemma A4 shows that there exists a non-stochastic b_T such that, for η such that (A7) is satisfied:

$$\sup_{i\in\{1,...,N\}} \sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \Pr\left(\widehat{g}_i\left(\theta,\alpha\right)\neq g_i^0\right) \quad \leq \quad b_T = o(T^{-\delta}).$$

Hence we have, for all $\delta>0$ and with probability one:

$$\sup_{(\theta,\alpha)\in\mathcal{N}_\eta} \Pr\left(\sup_{i\in\{1,...,N\}} \left|\widehat{g}_i\left(\theta,\alpha\right)-g_i^0\right|>0\right) \quad \leq \quad Nb_T = o\left(NT^{-\delta}\right).$$

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

A.3 Proof of Corollary 1

We have:

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

which tends to $\mathcal{N}\left(0,\Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1}\right)$ by Assumption 3.a-3.c and the Crámer theorem. Result (20) then follows from the fact that $\sqrt{NT}\left(\widehat{\theta}-\widetilde{\theta}\right)=o_p(1)$ and the Mann-Wald lemma.

Next we have, for all (g,t):

$$\begin{aligned}\widetilde{\alpha}_{gt} \quad &= \quad \frac{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}\left(y_{it}-x_{it}'\widetilde{\theta}\right)}{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}}\\ &= \quad \alpha_{gt}^0+\left(\frac{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}x_{it}}{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}}\right)'\left(\theta^0-\widetilde{\theta}\right)+\frac{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}v_{it}}{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}}.\end{aligned}$$

Now, using Assumptions 1.b and 2.a as well as the above we have:

$$\left(\frac{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}x_{it}}{\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}}\right)'\left(\theta^0-\widetilde{\theta}\right)=O_p\left(\frac{1}{\sqrt{NT}}\right).$$

Hence:

$$\sqrt{N}\left(\widetilde{\alpha}_{gt}-\alpha_{gt}^0\right) \quad = \quad \frac{\frac{1}{\sqrt{N}}\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}v_{it}}{\frac{1}{N}\sum_{i=1}^N\mathbf{1}\left\{g_i^0=g\right\}}+o_p(1),$$

and (21) follows from a similar argument as before.

This ends the proof of Corollary 1.

A.4 Proof of Proposition 1

Let $\overline{\theta}=\text{plim}_{N\to\infty}\widehat{\theta}$, and $\overline{\alpha}_g=\text{plim}_{N\to\infty}\widehat{\alpha}_g$ for $g\in\{1,2\}$, where the probability limits are taken for fixed T as N tends to infinity. We assume without loss of generality that $\overline{\alpha}_1\leq\overline{\alpha}_2$.

Following the arguments in Pollard (1981), it can be shown that the pseudo-true values $\overline{\theta}$ and $\overline{\alpha}_g$ satisfy:

$$\begin{aligned}
\mathbb{E}\Bigg[\sum_{t=1}^{T} x_{it}\left(v_{it}+x_{it}'\left(\theta^0-\overline{\theta}\right)\right)+\sum_{t=1}^{T} x_{it}\mathbf{1}\left\{\overline{v}_i\leq \overline{x}_i'\left(\overline{\theta}-\theta^0\right)+\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}-\alpha^0\right\}\left(\alpha^0-\overline{\alpha}_1\right)& \\
+\sum_{t=1}^{T} x_{it}\mathbf{1}\left\{\overline{v}_i> \overline{x}_i'\left(\overline{\theta}-\theta^0\right)+\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}-\alpha^0\right\}\left(\alpha^0-\overline{\alpha}_2\right)\Bigg] &= 0, \quad \text{(A19)}\\
\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i\leq \overline{x}_i'\left(\overline{\theta}-\theta^0\right)+\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}-\alpha^0\right\}\left(\overline{v}_i+\overline{x}_i'\left(\theta^0-\overline{\theta}\right)+\alpha^0-\overline{\alpha}_1\right)\right] &= 0, \quad \text{(A20)}\\
\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i> \overline{x}_i'\left(\overline{\theta}-\theta^0\right)+\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}-\alpha^0\right\}\left(\overline{v}_i+\overline{x}_i'\left(\theta^0-\overline{\theta}\right)+\alpha^0-\overline{\alpha}_2\right)\right] &= 0. \quad \text{(A21)}
\end{aligned}$$

Now, let a_1 and a_2 be the solutions of:

$$\begin{aligned}
T\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i\leq \frac{a_1+a_2}{2}-\alpha^0\right\}\left(\overline{v}_i+\alpha^0-a_1\right)\right] &= 0, \quad \text{(A22)}\\
T\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i> \frac{a_1+a_2}{2}-\alpha^0\right\}\left(\overline{v}_i+\alpha^0-a_2\right)\right] &= 0. \quad \text{(A23)}
\end{aligned}$$

Note that $\left(\theta^0,a_1,a_2\right)$ satisfies the moment restrictions (A19)-(A21) because, as v_{it} and x_{it} are independent of each other we have:

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T} x_{it}v_{it}+\sum_{t=1}^{T} x_{it}\mathbf{1}\left\{\overline{v}_i\leq \frac{a_1+a_2}{2}-\alpha^0\right\}\left(\alpha^0-a_1\right)+\sum_{t=1}^{T} x_{it}\mathbf{1}\left\{\overline{v}_i> \frac{a_1+a_2}{2}-\alpha^0\right\}\left(\alpha^0-a_2\right)\right]\\
&= 0+\mathbb{E}\left[\sum_{t=1}^{T} x_{it}\right]\underbrace{\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i\leq \frac{a_1+a_2}{2}-\alpha^0\right\}\left(\alpha^0-a_1\right)+\mathbf{1}\left\{\overline{v}_i> \frac{a_1+a_2}{2}-\alpha^0\right\}\left(\alpha^0-a_2\right)\right]}_{=0},
\end{aligned}$$

where we have used that the sum of the left-hand sides in (A22) and (A23) is zero.

Provided the solution to the population moment restrictions (A19)-(A21) be unique,[56] it thus follows that:

$$\left(\overline{\theta},\overline{\alpha}_1,\overline{\alpha}_2\right)=\left(\theta^0,a_1,a_2\right). \quad \text{(A24)}$$

Hence $\widehat{\theta}\xrightarrow{p}\theta^0$. In addition, it follows from (A22)-(A23) and (A24) that:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i\leq \frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}-\alpha^0\right\}\left(\overline{v}_i+\alpha^0-\overline{\alpha}_1\right)\right] &= 0,\\
\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i> \frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}-\alpha^0\right\}\left(\overline{v}_i+\alpha^0-\overline{\alpha}_2\right)\right] &= 0.
\end{aligned}$$

In particular we have, by symmetry: $(\overline{\alpha}_1+\overline{\alpha}_2)/2=\alpha^0$. So:

$$\overline{\alpha}_1=\alpha^0-\mathbb{E}\left(\overline{v}_i|\,\overline{v}_i\leq 0\right),$$

and likewise for $\overline{\alpha}_2$. The final result comes from the normality assumption, as:

$$\mathbb{E}\left(\overline{v}_i|\,\overline{v}_i\leq 0\right)=-\frac{\sigma}{\sqrt{T}}\frac{\phi(0)}{\Phi(0)}=-\sigma\sqrt{\frac{2}{\pi T}}.$$

This ends the proof of Proposition 1.

[56] Uniqueness of the population minimum is a key ingredient for showing that $\left(\widehat{\theta},\widehat{\alpha}\right)\xrightarrow{p}\left(\overline{\theta},\overline{\alpha}\right)$ as N tends to infinity (Pollard, 1981). Uniqueness is implicitly assumed in the statement of Proposition 1. See the supplementary appendix for details.

B Additional tables and figures

Table B1: Income and democracy, OLS and FE

	Unbalanced panel		Balanced panel	
	(1)	(2)	(3)	(4)
Lag democracy (θ_1)	.706 (.035)	.379 (.051)	.665 (.049)	.283 (.058)
Income (θ_2)	.072 (.010)	.010 (.035)	.083 (.014)	$-.031$ (.049)
Cumulative income ($\frac{\theta_2}{1-\theta_1}$)	.246 (.031)	.017 (.056)	.246 (.019)	$-.044$ (.069)
Observations	945	945	630	630
Countries	150	150	90	90
R-squared	.725	.796	.721	.799
Time dummies	yes	yes	yes	yes
Country fixed effects	no	yes	no	yes

Note: Balanced (1970-2000) and unbalanced (1960-2000) five-year panel data from Acemoglu et al. (2008). Freedom House indicator of democracy. Robust standard errors clustered at the country level in parentheses.

Table B2: Income and democracy, GFE estimates

G	Objective	BIC	Lag. dem.	Income	Cum. income
			(θ_1)	(θ_2)	$(\frac{\theta_2}{1-\theta_1})$
1	24.301	.052	.665	.083	.247
			(.049)	(.014)	(.018)
2	19.847	.046	.601	.061	.152
			(.041, .061, .072)	(.011, .013, .019)	(.021, .030, .058)
3	16.599	.042	.407	.089	.151
			(.052, .083, .129)	(.011, .015, .019)	(.013, .022, .036)
4	14.319	.039	.302	.082	.118
			(.054, .108, .140)	(.009, .012, .017)	(.011, .021, .038)
5	12.593	.037	.255	.079	.107
			(.050, .088, .134)	(.010, .012, .015)	(.009, .014, .040)
6	11.132	.036	.465	.064	.119
			(.043, .054, .122)	(.007, .008, .012)	(.011, .014.030)
7	10.059	.035	.403	.065	.108
			(.043, .074, .117)	(.008, .013, .013)	(.011, .019, .027)
8	9.251	.035	.333	.070	.104
			(.044, .085, .122)	(.008, .012, .013)	(.010, .014, .033)
9	8.426	.034	.312	.069	.101
			(.045, .072, .123)	(.008, .010, .013)	(.010, .011, .031)
10*	7.749	.034	.277	.075	.104
			(.049, .062, .124)	(.008, .010, .015)	(.009, .011, .034)
11	7.218	.034	.293	.073	.104
			(.042, .062, .130)	(.008, .012, .014)	(.009, .013, .030)
12	6.809	.034	.304	.074	.107
			(.044, .054, .109)	(.008, .009, .015)	(.009, .010, .037)
13	6.391	.035	.236	.072	.094
			(.040, .046, .120)	(.009, .010, .014)	(.009, .010, .031)
14	5.996	.035	.237	.071	.094
			(.042, .047, .119)	(.009, .010, .017)	(.009, .010, .038)
15	5.664	.035	.244	.071	.094
			(.043, .046, .127)	(.009, .010, .015)	(.009, .010, .040)
FE	17.517	–	.284	−.031	−.044
			(.058)	(.049)	(.069)

Note: See the notes to Figure 1. The table reports the value of the objective function, the Bayesian information criterion, and GFE coefficient estimates with their standard errors for various values of the number of groups G. Three different standard error estimates are shown in parentheses: based on the large-T normal approximation, on Pollard (1982)'s fixed-T normal approximation, and on the bootstrap, respectively. The parameter $\widehat{\sigma}^2$ in BIC was computed using $G_{max} = 15$. The last row in the table shows the same figures for fixed-effects regression.

Table B3: Income and democracy, GFE estimates with country-specific FE

G	Objective	Lag. dem.	Income	Cum. income
		(θ_1)	(θ_2)	$(\frac{\theta_2}{1-\theta_1})$
1	17.517	.284 (.058)	$-.031$ (.049)	$-.044$ (.069)
2	12.859	.061 (.049)	$-.038$ (.027)	$-.040$ (.029)
3	10.400	$-.033$ (.043)	$-.035$ (.027)	$-.034$ (.027)
4	9.221	$-.072$ (.046)	.045 (.027)	.042 (.025)
5	8.174	$-.093$ (.042)	$-.013$ (.026)	$-.011$ (.024)

Note: See the notes to Table B2. The table reports GFE estimates in deviations to country-specific means (i.e., net of country FE). Clustered standard errors based on the large-T normal approximation in parentheses.

Table B4: Binary measure of democracy, transitions 1970/2000 by group ($G = 4$)

Transition 1970/2000	0/0	0/1	1/0	1/1	All
Group 1 (“high-democracy”)	0	0	3	30	33
Group 2 (“low-democracy”)	25	1	0	0	26
Group 3 (“early transition”)	0	12	0	1	13
Group 4 (“late transition”)	3	12	1	2	18
All	28	25	4	33	90

Note: See the notes to Table B2. Here we code as “non-democratic” (that is, 0) countries whose Freedom House score is lower than .50, and as “democratic” (that is, 1) countries with a score $> .50$. The numbers a/b in the table denote transition from state $a \in \{0, 1\}$ in 1970 to state $b \in \{0, 1\}$ in 2000.

Table B5: Descriptive statistics, by group

Group	1 (high dem.)	2 (low dem.)	3 (early trans.)	4 (late trans.)
log GDP p.c. (1500)	6.52 (.300)	6.39 (.437)	6.49 (.141)	6.30 (.236)
Independence Year	1860 (63.3)	1939 (50.7)	1824 (37.7)	1924 (56.3)
Constraints	.581 (.446)	.258 (.254)	.125 (.166)	.250 (.246)
Democracy (1965)	.892 (.157)	.446 (.171)	.510 (.267)	.508 (.281)
log GDP p.c. (1965)	8.76 (.765)	7.33 (.604)	8.02 (.709)	7.39 (.773)
Education (1970)	5.78 (2.59)	1.52 (1.05)	3.63 (1.61)	2.59 (1.92)
Share Catholic (1980)	.434 (.404)	.232 (.284)	.626 (.437)	.379 (.349)
Share Protestant (1980)	.248 (.330)	.068 (.088)	.024 (.032)	.140 (.160)
Number of observations	33	26	13	18

Note: Balanced panel from Acemoglu et al. (2008). "Constraints" are constraints on the executive at independence, measured as in Acemoglu et al. (2005). Group-specific means, and group-specific standard deviations in parentheses. Group membership is shown on Figure 2.

Table B6: Group membership estimates, various specifications

Country	$G = 2$	$G = 3$	$G = 4$	$G = 5$	$G = 6$	Two-layer		FE ($G = 3$)
Algeria	2	2	2	2	2	Stable	Low	Stable
Argentina	1	3	3	3	3	Early	Low	Early
Australia	1	1	1	1	1	Stable	High	Stable
Austria	1	1	1	1	1	Stable	High	Stable
Belgium	1	1	1	1	1	Stable	High	Stable
Benin	2	3	4	4	4	Late	Low	Late
Bolivia	1	3	3	3	3	Early	Low	Late
Brazil	1	3	3	3	3	Early	Low	Early
Burkina Faso	1	1	4	5	5	Stable	Medium-Low	Stable
Burundi	2	2	2	2	2	Stable	Low	Stable
Cameroon	2	2	2	2	2	Stable	Low	Stable
Canada	1	1	1	1	1	Stable	High	Stable
Central African Rep.	2	3	4	4	4	Late	Low	Late
Chad	2	2	2	2	2	Stable	Low	Stable
Chile	1	3	4	5	5	Late	High	Late
China	2	2	2	2	2	Stable	Low	Stable
Colombia	1	1	1	1	1	Stable	Medium-High	Stable
Congo, Dem. Rep.	2	2	2	2	2	Stable	Low	Stable
Congo Republic	2	2	2	2	2	Stable	Low	Stable
Costa Rica	1	1	1	1	1	Stable	High	Stable
Cote d'Ivoire	2	2	2	2	2	Stable	Low	Stable
Cyprus	1	1	1	1	1	Stable	Medium-High	Late
Denmark	1	1	1	1	1	Stable	High	Stable
Dominican Republic	1	1	1	1	1	Stable	Medium-High	Stable
Ecuador	2	3	3	3	6	Early	Low	Early
Egypt	2	2	2	2	2	Stable	Medium-Low	Stable
El Salvador	1	1	1	1	3	Stable	Medium-High	Stable
Finland	1	1	1	1	1	Stable	Medium-High	Stable
France	1	1	1	1	1	Stable	High	Stable
Gabon	2	2	2	2	2	Stable	Low	Stable
Ghana	2	3	4	4	6	Late	High	Late
Greece	2	3	3	3	3	Early	High	Early
Guatemala	1	1	1	5	5	Stable	Medium	Stable
Guinea	2	2	2	2	2	Stable	Low	Stable
Honduras	2	3	3	3	3	Early	Low	Early

Table B6: Group membership estimates, various specifications (cont.)

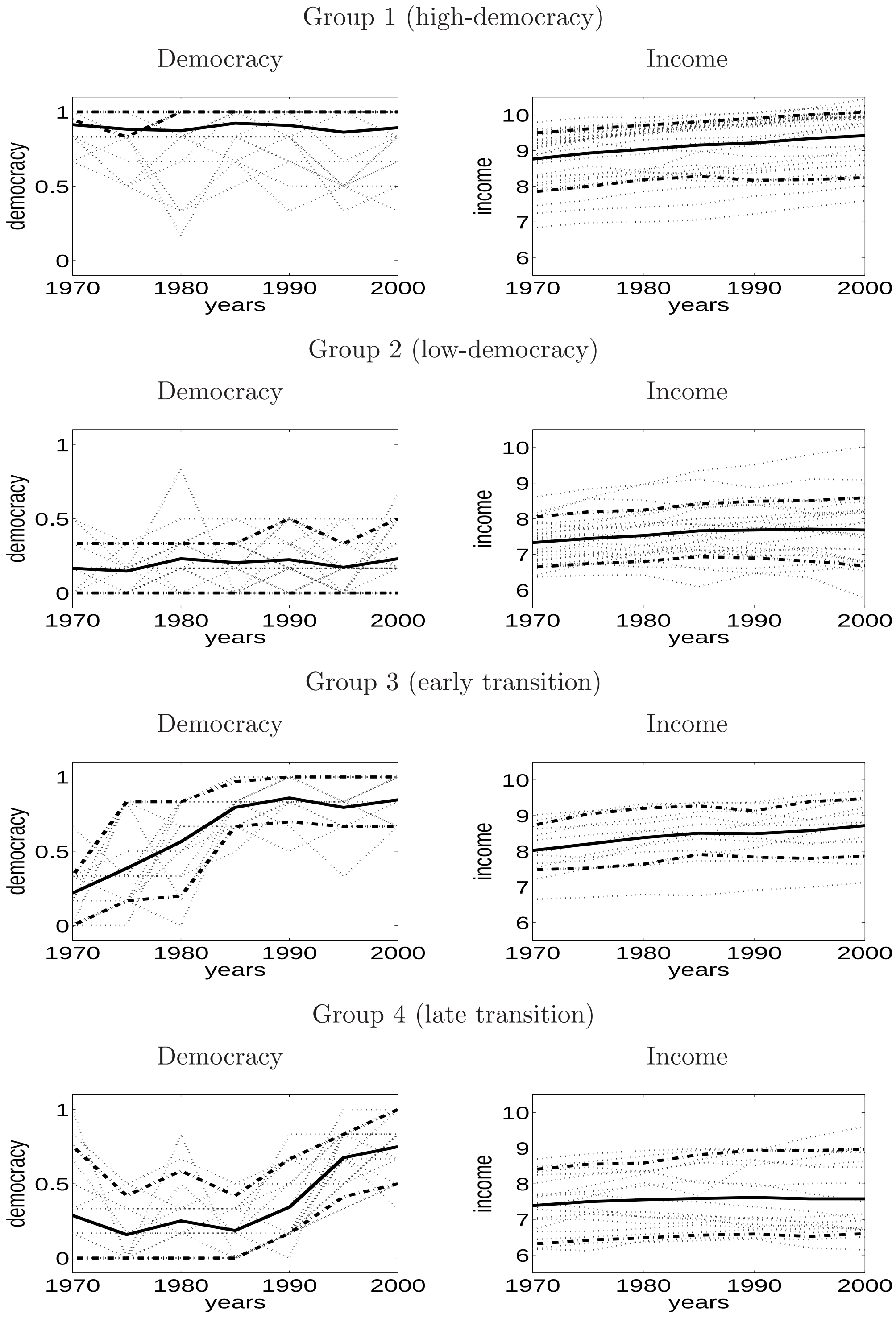
Country	$G = 2$	$G = 3$	$G = 4$	$G = 5$	$G = 6$	Two-layer		FE ($G = 3$)
Iceland	1	1	1	1	1	Stable	High	Stable
India	1	1	1	1	1	Stable	High	Stable
Indonesia	1	2	2	5	5	Stable	Medium-Low	Stable
Iran	2	2	2	2	2	Stable	Low	Stable
Ireland	1	1	1	1	1	Stable	High	Stable
Israel	1	1	1	1	1	Stable	Medium-High	Stable
Italy	1	1	1	1	1	Stable	High	Stable
Jamaica	1	1	1	1	1	Stable	High	Stable
Japan	1	1	1	1	1	Stable	High	Stable
Jordan	2	2	2	2	2	Stable	Medium-Low	Late
Kenya	2	2	2	2	2	Stable	Medium-Low	Stable
Korea, Rep.	1	3	3	3	3	Early	Low	Late
Luxembourg	1	1	1	1	1	Stable	High	Stable
Madagascar	2	3	4	4	4	Late	High	Late
Malawi	2	3	4	4	4	Late	Low	Late
Malaysia	1	1	1	5	1	Stable	Medium	Stable
Mali	2	3	4	4	4	Late	Low	Late
Mauritania	2	2	2	2	2	Stable	Low	Stable
Mexico	2	2	4	5	6	Stable	Medium	Stable
Morocco	1	2	2	5	2	Stable	Medium-Low	Stable
Nepal	1	1	3	3	1	Early	Low	Early
Netherlands	1	1	1	1	1	Stable	High	Stable
New Zealand	1	1	1	1	1	Stable	High	Stable
Nicaragua	1	3	4	5	5	Stable	Medium	Stable
Niger	2	3	4	4	4	Late	Low	Late
Nigeria	2	2	2	5	6	Stable	Medium-Low	Stable
Norway	1	1	1	1	1	Stable	High	Stable
Panama	2	3	4	4	6	Late	Low	Late
Paraguay	1	2	2	5	5	Stable	Medium-Low	Stable
Peru	2	2	3	3	6	Early	Low	Early
Philippines	2	3	4	3	4	Late	High	Late
Portugal	1	1	3	3	1	Early	High	Early
Romania	2	3	4	4	4	Late	Low	Late
Rwanda	2	2	2	2	2	Stable	Low	Stable
Sierra Leone	2	2	2	5	5	Stable	Medium-Low	Stable

Table B6: Group membership estimates, various specifications (cont.)

Country	$G = 2$	$G = 3$	$G = 4$	$G = 5$	$G = 6$	Two-layer		FE ($G = 3$)
Singapore	2	2	2	2	2	Stable	Low	Stable
South Africa	1	3	4	4	4	Late	High	Late
Spain	1	1	3	3	1	Early	High	Early
Sri Lanka	1	1	1	1	1	Stable	Medium-High	Stable
Sweden	1	1	1	1	1	Stable	High	Stable
Switzerland	1	1	1	1	1	Stable	High	Stable
Syria	2	2	2	2	2	Stable	Low	Stable
Taiwan	2	3	4	4	5	Late	High	Late
Tanzania	2	3	4	4	4	Stable	Medium-Low	Stable
Thailand	1	1	3	3	3	Early	High	Early
Togo	2	2	2	2	2	Stable	Low	Stable
Trinidad and Tobago	1	1	1	1	1	Stable	High	Stable
Tunisia	2	2	2	2	2	Stable	Low	Stable
Uganda	2	2	2	2	2	Stable	Medium-Low	Stable
United Kingdom	1	1	1	1	1	Stable	High	Stable
United States	1	1	1	1	1	Stable	High	Stable
Uruguay	1	3	3	3	3	Early	High	Late
Venezuela	1	1	1	1	1	Stable	Medium-High	Stable
Zambia	2	3	4	4	4	Stable	Medium-Low	Stable

Note: Group membership, on the balanced panel from Acemoglu et al. (2008). Columns 2 to 6 show the GFE estimates, for $G = 2, ..., 6$. The next two columns show estimates from a two-layer specification, with $G_1 = 3$ ("Stable", "Early", and "Late", respectively), and $G_2 = \{5, 2, 2\}$ ("High" and "Low", with "Medium-High", "Medium" and "Medium-Low" as intermediate categories for stable countries). The last column shows GFE estimates in deviations to country-specific means, for $G = 3$.

Figure B1: Confidence bands and data paths of democracy and income ($G = 4$)


Group 1 (high-democracy)

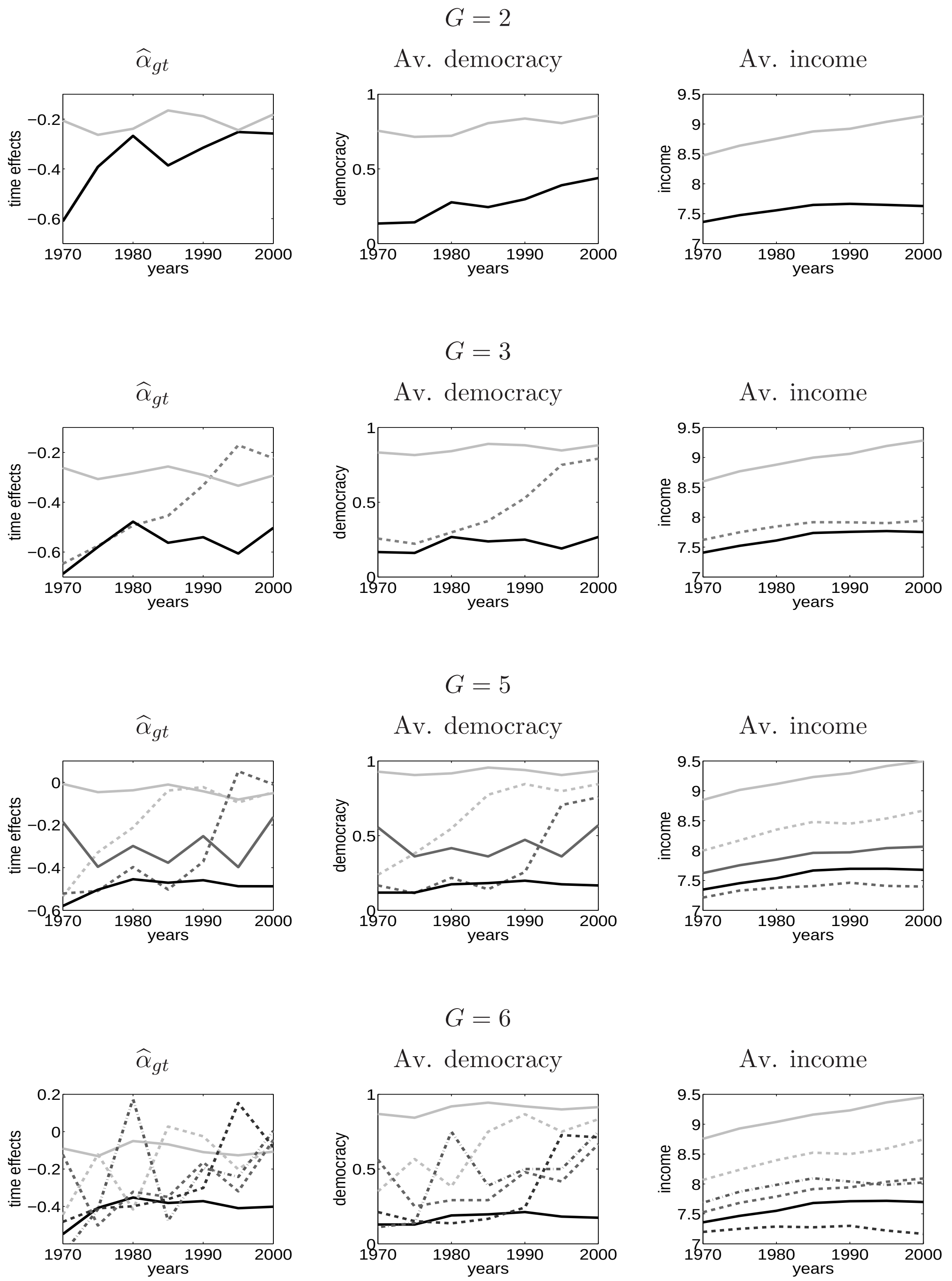
Group 2 (low-democracy)

Group 3 (early transition)

Group 4 (late transition)

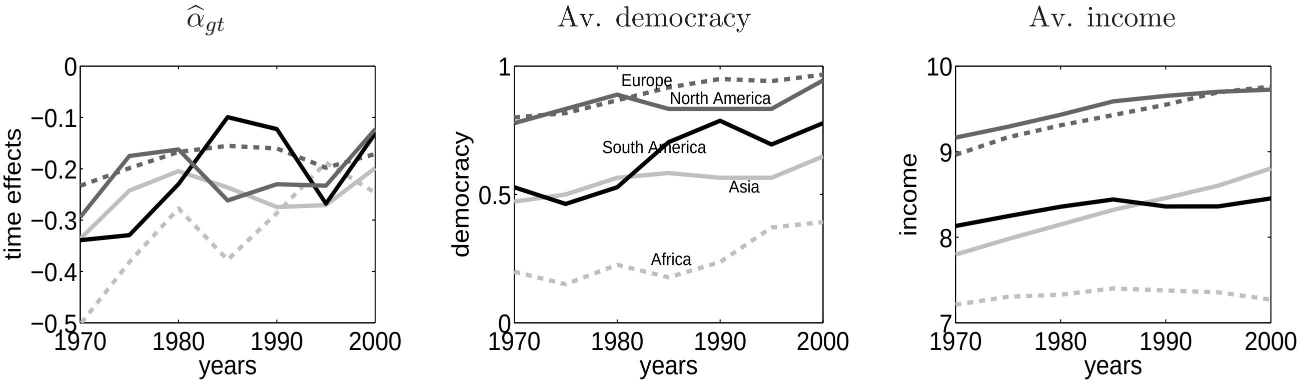
Note: See the notes to Figure 1. The left column shows the mean normalized Freedom House score (thick solid lines), a uniform 50%-confidence band (thick dashed-dotted lines), as well as the plot of all democracy paths in the data (thin dotted lines), by group. The right column shows the same figures for lagged log-GDP per capita.

Figure B2: Patterns of heterogeneity, various G



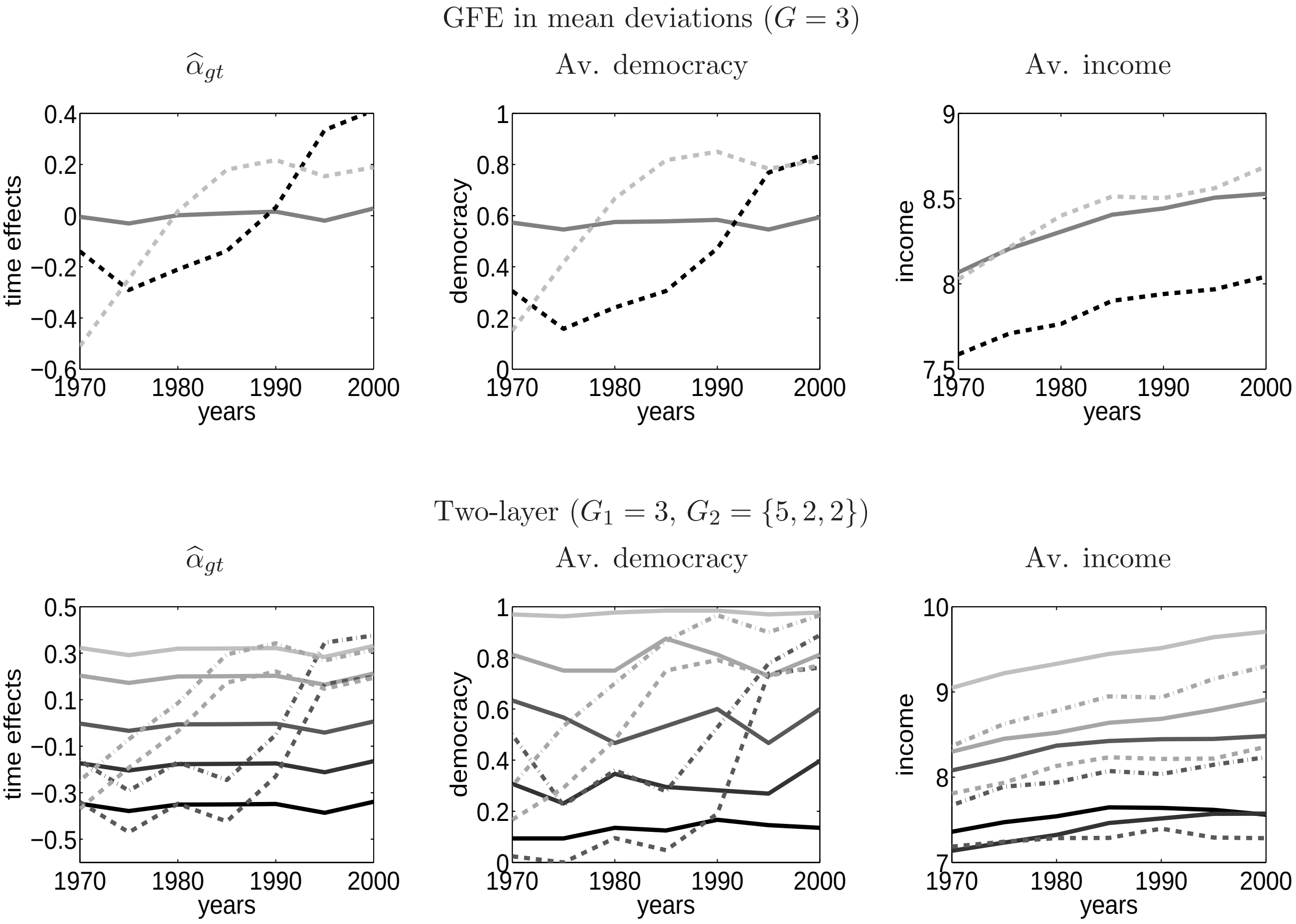
Note: See the notes to Figure 1. The left column reports the group-specific time effects $\widehat{\alpha}_{gt}$ for $G = 2$, $G = 3$, $G = 5$, and $G = 6$, from top to bottom. The other two columns show the group-specific averages of democracy and lagged log-GDP per capita, respectively. Calendar years $(1970 - 2000)$ are shown on the x-axis.

Figure B3: Continent-specific time-effects



Note: See the notes to Figure B2. The five groups are Europe, North-America (including Mexico), South-America, Asia (including Australia and New-Zealand), and Africa.

Figure B4: Grouped patterns, alternative specifications


GFE in mean deviations ($G = 3$)

Two-layer ($G_1 = 3$, $G_2 = \{5, 2, 2\}$)

Note: See the notes to Figure B2. The top panel shows the results of GFE estimation in deviation to country-specific means. The bottom panel shows the results of the two-layer specification (7).

Supplementary Appendix to
"Grouped Patterns of Heterogeneity in Panel Data"

Stéphane Bonhomme and Elena Manresa

This supplementary appendix contains a study of fixed-T inference for the grouped fixed-effects (GFE) estimator. It also considers several issues discussed in the main text. Finally, it describes additional results relative to the Monte Carlo exercise.

S1 Fixed-T inference

In this section of the supplementary appendix we study the fixed-T asymptotic properties of the GFE estimator.

S1.1 Asymptotic distribution

Let $\left(\widehat{\theta},\widehat{\alpha}\right)$ be the GFE estimator of (θ,α). Let also $y_i=(y_{i1},...,y_{iT})'$ (with dimensions $T\times 1$), and $x_i=(x_{i1},...,x_{iT})'$ ($T\times K$, where $K=\dim x_{it}$). We assume that (y_i,x_i) are i.i.d. across individuals and have finite second moments. In addition, we assume that the solution to the following population minimization problem:

$$\left(\overline{\theta},\overline{\alpha}\right)=\underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmin}}\ \mathbb{E}\left[\sum_{t=1}^{T}\left(y_{it}-x_{it}'\theta-\alpha_{\widehat{g}_i(\theta,\alpha)t}\right)^2\right], \tag{S1}$$

is unique up to relabelling. Lastly, we assume that the solution to every minimization problem of the form (S1) but based on $\widetilde{G}<G$ groups is also unique. Then, extending the analysis of Pollard (1981) to allow for covariates, it can be shown that, as N tends to infinity with T fixed:

$$\left(\widehat{\theta},\widehat{\alpha}\right)\xrightarrow{p}\left(\overline{\theta},\overline{\alpha}\right).$$

Note that, in contrast with the asymptotic analysis of Section 4, uniqueness of the solution in (S1) does not require the data generating process to have a group structure.[1]

If the conditions of Pollard (1981)'s consistency theorem are satisfied, the pseudo-true parameter value $\left(\overline{\theta},\overline{\alpha}\right)$ solves the following system of moment restrictions:

$$\mathbb{E}\left[x_i'\left(y_i-x_i\overline{\theta}-\overline{\alpha}_{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)}\right)\right]=0, \tag{S2}$$

and:

$$\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\left(y_i-x_i\overline{\theta}-\overline{\alpha}_g\right)\right]=0, \quad \text{for all } g=1,...,G, \tag{S3}$$

where $\overline{\alpha}_g=(\overline{\alpha}_{g1},...,\overline{\alpha}_{gT})'$ is $T\times 1$. As in the main text we will also denote as $\alpha=(\alpha_1',\alpha_2',...,\alpha_G')'$ the $GT\times 1$ vector that stacks all α_{gt}'s.

[1] On the other hand, this assumption rules out purely homogeneous DGPs as soon as $T\geq 2$. To see this, suppose that y_{it} are i.i.d. standard normal, and that there are no covariates in the model. In the case $T=2$, it can be shown that the solutions to (S1) lie on a circle whose radius is identified, but that the precise location of the points on the circle is not.

Using empirical process theory, Pollard (1982) shows that, in the absence of covariates, $\sqrt{N}\left(\widehat{\alpha}-\overline{\alpha}\right)$ is asymptotically normally distributed under suitable conditions. Adapting Pollard's arguments to account for covariates, it can be shown that:

$$\sqrt{N}\left(\begin{array}{c}\widehat{\theta}-\overline{\theta}\\ \widehat{\alpha}-\overline{\alpha}\end{array}\right)\overset{d}{\to}\mathcal{N}\left(0,\Gamma^{-1}V\Gamma^{-1}\right), \tag{S4}$$

where the $(GT+K)\times(GT+K)$ matrices V and Γ are defined below. As in Pollard (1982)'s main theorem, for (S4) to hold we assume that y_i has a continuous density given x_i, and that Γ is positive definite, in addition to the assumptions needed for consistency.

Note that the GFE estimator $\left(\widehat{\theta},\widehat{\alpha}\right)$ is a just-identified GMM estimator based on non-smooth moment functions. As a consequence, V is given by:

$$V=\mathbb{E}\left[W_i\left(\overline{\theta},\overline{\alpha}\right)\left(y_i-x_i\overline{\theta}-\overline{\alpha}_{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)}\right)\left(y_i-x_i\overline{\theta}-\overline{\alpha}_{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)}\right)'W_i\left(\overline{\theta},\overline{\alpha}\right)'\right],$$

where:

$$W_i\left(\overline{\theta},\overline{\alpha}\right)=\left(\begin{array}{c}x_i'\\ e_{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)}\otimes I_T\end{array}\right),$$

and where $e_1,...,e_G$ denotes the canonical basis of $\mathbb{R}^G$.

Moreover, Γ is given by:

$$\Gamma=\left(\begin{array}{cccc}\Gamma_{\theta\theta} & \Gamma_{\theta 1} & ... & \Gamma_{\theta G}\\ \Gamma_{1\theta} & \Gamma_{11} & ... & \Gamma_{1G}\\ ... & ... & ... & ...\\ \Gamma_{G\theta} & \Gamma_{G1} & ... & \Gamma_{GG}\end{array}\right),$$

where:

$$\Gamma_{\theta\theta}=-\frac{\partial}{\partial\theta'}\Big|_{\left(\overline{\theta},\overline{\alpha}\right)}\mathbb{E}\left[x_i'\left(y_i-x_i\theta-\alpha_{\widehat{g}_i(\theta,\alpha)}\right)\right],$$

$$\Gamma_{\theta g}=-\frac{\partial}{\partial\alpha_g'}\Big|_{\left(\overline{\theta},\overline{\alpha}\right)}\mathbb{E}\left[x_i'\left(y_i-x_i\theta-\alpha_{\widehat{g}_i(\theta,\alpha)}\right)\right],$$

$$\Gamma_{g\widetilde{g}}=-\frac{\partial}{\partial\alpha_{\widetilde{g}}'}\Big|_{\left(\overline{\theta},\overline{\alpha}\right)}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\left(y_i-x_i\theta-\alpha_g\right)\right],$$

and where $\Gamma_{g\theta}=\Gamma_{\theta g}'$.

The next result provides a convenient alternative expression for Γ.

Proposition S1 *Let us denote as f the conditional density of y_i given x_i. Let us also define, for all $(g,h)\in\{1,...,G\}^2$:*

$$\begin{aligned}S_{gh} \quad = \quad & \Big\{y\in\mathbb{R}^T, \|y-x\theta-\alpha_g\|^2=\|y-x\theta-\alpha_h\|^2, \text{ and}\\ & \|y-x\theta-\alpha_g\|^2\leq\|y-x\theta-\alpha_{\widetilde{h}}\|^2 \text{ for all } \widetilde{h}\neq(g,h)\Big\}.\end{aligned} \tag{S5}$$

We denote S_{gh} as $\overline{S}_{gh}$ when evaluated at $\left(\overline{\theta},\overline{\alpha}\right)$.[2]

[2] Note that S_{gh} and $\overline{S}_{gh}$ depend on x, although we leave the dependence implicit for conciseness.

We have:

$$\Gamma_{\theta\theta} \;=\; \mathbb{E}\left[x_i'x_i\right] - \frac{1}{2}\sum_{g=1}^{G}\sum_{h\neq g}\mathbb{E}\left[\left(\int_{\overline{S}_{gh}} f(y|x_i)dy\right) x_i'\left(\frac{(\overline{\alpha}_h-\overline{\alpha}_g)(\overline{\alpha}_h-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}\right)x_i\right], \tag{S6}$$

$$\Gamma_{\theta g} \;=\; \mathbb{E}\left[x_i'\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\right] + \sum_{h\neq g}\mathbb{E}\left[x_i'\left(\overline{\alpha}_g-\overline{\alpha}_h\right)\left(\int_{\overline{S}_{gh}}\frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}f(y|x_i)dy\right)\right], \tag{S7}$$

$$\Gamma_{gg} \;=\; \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\right]I_T - \mathbb{E}\left[\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}\frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)(y-x_i\overline{\theta}-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}f(y|x_i)dy\right)\right], \tag{S8}$$

$$\Gamma_{g\widetilde{g}} \;=\; \mathbb{E}\left[\left(\int_{\overline{S}_{g\widetilde{g}}}\frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)(y-x_i\overline{\theta}-\overline{\alpha}_{\widetilde{g}})'}{\|\overline{\alpha}_{\widetilde{g}}-\overline{\alpha}_g\|}f(y|x_i)dy\right)\right],\ \textit{for all } \widetilde{g}\neq g. \tag{S9}$$

Proof. See the appendix. ■

The regions $\overline{S}_{gh}$ comprise units that are at the margin between belonging to groups g or h. The large-T variance is obtained when f has no mass on $\overline{S}_{gh}$. In a fixed-T asymptotic, in contrast, group misclassification adds an extra contribution to the variance of the GFE estimator.

Example. Consider the simple case with no covariates, time-invariant heterogeneity $\alpha_{g_it}=\alpha_{g_i}$, and $G=2$. In this case the pseudo-true value $(\overline{\alpha}_1,\overline{\alpha}_2)$ satisfies:

$$\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\overline{\alpha}_1,\overline{\alpha}_2\right)=g\right\}\left(\overline{y}_i-\overline{\alpha}_g\right)\right]=0,\quad g=1,2.$$

That is, assuming $\overline{\alpha}_1<\overline{\alpha}_2$ without loss of generality:

$$\int_{-\infty}^{\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}}\left(y-\overline{\alpha}_1\right)f(y)dy=0,\qquad\text{and}\qquad\int_{\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}}^{+\infty}\left(y-\overline{\alpha}_2\right)f(y)dy=0,$$

where $f(y)$ denotes the density of $\overline{y}_i$.

It is easily verified that:

$$\Gamma=\begin{pmatrix}\mathbb{E}\left(\mathbf{1}\left\{\widehat{g}_i\left(\overline{\alpha}_1,\overline{\alpha}_2\right)=1\right\}\right) & 0\\ 0 & \mathbb{E}\left(\mathbf{1}\left\{\widehat{g}_i\left(\overline{\alpha}_1,\overline{\alpha}_2\right)=2\right\}\right)\end{pmatrix}-\left|\frac{\overline{\alpha}_2-\overline{\alpha}_1}{4}\right|f\left(\frac{\overline{\alpha}_1+\overline{\alpha}_2}{2}\right)\begin{pmatrix}1&1\\1&1\end{pmatrix}.$$

The second term in Γ represents the contribution to the variance due to observations that are at the margin between group 1 and group 2.

S1.2 Variance estimation

We study two strategies in turn: variance estimation based on analytical formulas, and inference based on the bootstrap.

Analytical formulas. A consistent estimator of V is readily obtained as:

$$\widehat{V}=\frac{1}{N}\sum_{i=1}^{N}W_i\left(\widehat{\theta},\widehat{\alpha}\right)\left(y_i-x_i\widehat{\theta}-\widehat{\alpha}_{\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)}\right)\left(y_i-x_i\widehat{\theta}-\widehat{\alpha}_{\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right)}\right)'W_i\left(\widehat{\theta},\widehat{\alpha}\right)'.$$

To construct a consistent estimator of Γ, we use the following:

$$\widehat{\Gamma}_{\theta\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i' x_i - \frac{1}{2N}\sum_{g=1}^{G}\sum_{h\neq g}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) x_i' \left(\frac{(\widehat{\alpha}_h-\widehat{\alpha}_g)(\widehat{\alpha}_h-\widehat{\alpha}_g)'}{\|\widehat{\alpha}_h-\widehat{\alpha}_g\|}\right) x_i, \tag{S10}$$

$$\widehat{\Gamma}_{\theta g} = \frac{1}{N}\sum_{i=1}^{N} x_i' \mathbf{1}\left\{\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right) = g\right\} + \frac{1}{N}\sum_{h\neq g}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) x_i' \left(\widehat{\alpha}_g-\widehat{\alpha}_h\right) \frac{(y_i - x_i\widehat{\theta}-\widehat{\alpha}_g)'}{\|\widehat{\alpha}_h-\widehat{\alpha}_g\|}, \tag{S11}$$

$$\widehat{\Gamma}_{gg} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left\{\widehat{g}_i\left(\widehat{\theta},\widehat{\alpha}\right) = g\right\} I_T - \frac{1}{N}\sum_{h\neq g}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) \frac{(y_i - x_i\widehat{\theta}-\widehat{\alpha}_g)(y_i - x_i\widehat{\theta}-\widehat{\alpha}_g)'}{\|\widehat{\alpha}_h-\widehat{\alpha}_g\|}, \tag{S12}$$

$$\widehat{\Gamma}_{g\widetilde{g}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) \frac{(y_i - x_i\widehat{\theta}-\widehat{\alpha}_g)(y_i - x_i\widehat{\theta}-\widehat{\alpha}_{\widetilde{g}})'}{\|\widehat{\alpha}_{\widetilde{g}}-\widehat{\alpha}_g\|}, \text{ for all } \widetilde{g}\neq g, \tag{S13}$$

where:

$$\begin{aligned}\widehat{\Delta}_{igh}(\epsilon_N) &= \frac{1}{\epsilon_N}\kappa\left(\frac{\left(\frac{\widehat{\alpha}_h-\widehat{\alpha}_g}{\|\widehat{\alpha}_h-\widehat{\alpha}_g\|}\right)'\left(y_i - x_i\widehat{\theta} - \frac{\widehat{\alpha}_g+\widehat{\alpha}_h}{2}\right)}{\epsilon_N}\right)\\ &\quad \times \mathbf{1}\left\{\max\left(\left\|y_i - x_i\widehat{\theta}-\widehat{\alpha}_g\right\|^2, \left\|y_i - x_i\widehat{\theta}-\widehat{\alpha}_h\right\|^2\right) \leq \min_{\widetilde{h}\neq(g,h)} \left\|y_i - x_i\widehat{\theta}-\widehat{\alpha}_{\widetilde{h}}\right\|^2\right\},\end{aligned}$$

and where $\kappa(\cdot)$ is a kernel function. Note that $\frac{\overline{\alpha}_h-\overline{\alpha}_g}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}$ is the normal vector to the hypersurface $\overline{S}_{gh}$. The estimator $\widehat{\Gamma}$ is reminiscent of Powell (1986)'s variance estimator for quantile regression. Similarly, $\widehat{\Gamma}$ will be consistent for Γ if $\epsilon_N \to 0$ and $\sqrt{N}\epsilon_N \to +\infty$. To implement this method we take a Gaussian kernel $\kappa = \phi$. Optimal choice of ϵ_N exceeds the scope of this paper.[3]

Bootstrap. An alternative to the analytical formulas $\widehat{V}$ and $\widehat{\Gamma}$ is to use the bootstrap, resampling unit-specific blocks of observations (y_i, x_i) from the original sample. Consistency of the bootstrap for the minimum sum-of-squares partitioning problem, relying on the asymptotic derivations of Pollard (1982) and the results on the bootstrap obtained by Giné and Zinn (1990), is shown in Arcones and Giné (1992). As it requires multiple optimization of the GFE objective for different samples, however, the bootstrap is computationally intensive.

S2 Complements

In this section of the supplementary appendix we study three issues in turn: the link between GFE and finite mixtures, how to incorporate prior information to GFE, and estimation in unbalanced panels.

S2.1 A finite mixture interpretation

Here we show that the grouped fixed-effects estimator can be interpreted as the maximizer of the pseudo-likelihood of a mixture-of-normals model, where the mixing probabilities are individual-specific and unrestricted.

[3] We experimented with the following non-adaptive rule, roughly mimicking Silverman (1986)'s rule of thumb for density estimation:

$$\epsilon_N = 1.06 \min_{g,h\neq g}\left(\sqrt{\widehat{\text{Var}}\left(\left(\frac{\widehat{\alpha}_h-\widehat{\alpha}_g}{\|\widehat{\alpha}_h-\widehat{\alpha}_g\|}\right)'\left(y_i - x_i\widehat{\theta}\right)\right)}\right) N^{-\frac{1}{5}},$$

and obtained good results on simulated and real data. This is the choice we used in Tables 4, B2, and S2.

This contrasts with standard finite mixture modelling (McLachlan and Peel, 2000), which typically specifies the group probabilities $\pi_g(x_i)$ as functions of the covariates. In comparison, in the grouped fixed-effects approach the group probabilities $\pi_{ig} = \pi_g(i)$ are unrestricted functions of the individual dummies. We show the result in the case of the linear model (1), although the equivalence applies to nonlinear models also.

To state the equivalence result, let $\sigma > 0$ be a scaling parameter. Then, it is easy to see that the GFE estimator of (θ, α) satisfies:

$$\left(\widehat{\theta}, \widehat{\alpha}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmax}} \left[\max_{\pi_1,\ldots,\pi_N} \sum_{i=1}^{N} \ln \left(\sum_{g=1}^{G} \pi_{ig} \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{gt}\right)^2 \right)\right)\right], \tag{S14}$$

where the maximum is taken over all probability vectors $\pi_i = (\pi_{i1}, ...\pi_{iG})$ in the unit simplex of $\mathbb{R}^G$. The result comes from the fact that the individual-specific π_i are unrestricted.[4] Note also that (S14) holds for any choice of σ.

S2.2 Adding prior information

The GFE estimator may easily be modified to incorporate prior information on group membership. To proceed, suppose that prior information takes the form of probabilities, and denote as π_{ig} the prior probability that unit i belongs to group g. A penalized GFE estimator of (θ, α) is:

$$\left(\widehat{\theta}^{(\pi)}, \widehat{\alpha}^{(\pi)}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{\widehat{g}_i^{(\pi_i)}(\theta,\alpha)t}\right)^2, \tag{S15}$$

where the estimated group variables are now:

$$\widehat{g}_i^{(\pi_i)}(\theta, \alpha) = \underset{g\in\{1,\ldots,G\}}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{gt}\right)^2 - C \ln \pi_{ig}, \tag{S16}$$

and where $C > 0$ is a penalty term. The penalty specifies the respective weights attached to prior and data information in estimation.[5]

Note that computation of the penalized GFE estimator is very similar to that of the GFE estimator given by (4).[6] In addition, the penalized and unpenalized GFE estimators are asymptotically equivalent under the conditions given in Section 4, provided prior information is non-dogmatic in the following sense:

[4] Specifically, given (θ, α) values the maximum is achieved at:

$$\widehat{\pi}_i(\theta, \alpha) = \underset{\pi_i}{\operatorname{argmax}} \sum_{g=1}^{G} \pi_{ig} \frac{1}{(2\pi\sigma^2)^{\frac{T}{2}}} \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(y_{it} - x_{it}'\theta - \alpha_{gt}\right)^2 \right),$$

yielding:

$$\widehat{\pi}_{ig}(\theta, \alpha) = \mathbf{1}\left\{\widehat{g}_i(\theta, \alpha) = g\right\}, \quad \text{for all } g.$$

[5] A possible choice, motivated by the special case of the normal linear model, is $C = 2\sigma^2$, where $\sigma^2 = \mathbb{E}(v_{it}^2)$. In practice, one may approximate σ^2 by taking the mean of (OLS) squared residuals.

[6] We observed in numerical experiments that adding prior information tends to alleviate the local minima problem documented in Section 3, although it does not fully solve it.

Assumption S1 *(prior probabilities) The prior probabilities are non-dogmatic in the sense that, for some* $\varepsilon > 0$*:*

$$\varepsilon < \pi_{ig} < 1 - \varepsilon, \quad \textit{for all } (i, g).$$

We have the following result.

Corollary S1 *(penalized GFE) Let the assumptions of Corollary 1 hold, and let* $\pi = \{\pi_{ig}\}$ *be a set of prior probabilities that satisfies Assumption S1. Then we have, asymptotically:*

$$\sqrt{NT}\left(\widehat{\theta}^{(\pi)} - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1}\right). \tag{S17}$$

Proof. The proof closely follows that of Theorem 2 and Corollary 1. A difference appears in the proof of Lemma A4. Let us define the following quantity:

$$Z_{ig}^{(\pi)}(\theta, \alpha) = \mathbf{1}\{g_i^0 \neq g\}\mathbf{1}\left\{\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{gt}\right)^2 - C\ln\pi_{ig} \leq \sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{g_i^0 t}\right)^2 - C\ln\pi_{i,g_i^0}\right\}.$$

The proof consists in bounding $Z_{ig}^{(\pi)}(\theta, \alpha)$ instead of bounding $Z_{ig}(\theta, \alpha)$. The only difference is the following extra term in A_T:

$$A_{4T} = \left|-C\ln\pi_{ig} + C\ln\pi_{i\widetilde{g}}\right|,$$

which is bounded as follows:

$$A_{4T} \leq C\ln\left(\frac{1-\varepsilon}{\varepsilon}\right),$$

where we have used Assumption S1.

■

S2.3 Grouped fixed-effects in unbalanced panels

Let us consider an unbalanced panel whose maximum time length is T. We denote as d_{it} the indicator variable that takes value one if observations y_{it} and x_{it} belong to the dataset, and zero otherwise. We adopt the convention that $d_{it}y_{it} = 0$ and $d_{it}x_{it} = 0$ when the latter situation happens. It is assumed that x_{it} and v_{it} are weakly uncorrelated given $d_{it} = 1$.

The GFE estimator is then:

$$\left(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} d_{it}\left(y_{it} - x_{it}'\theta - \alpha_{g_i t}\right)^2. \tag{S18}$$

In terms of computation, one difference with Algorithm 1 arises in the update step, as it may happen that

$$n_{gt} = \sum_{i=1}^{N} d_{it}\mathbf{1}\left\{g_i^{(s+1)} = g\right\}$$

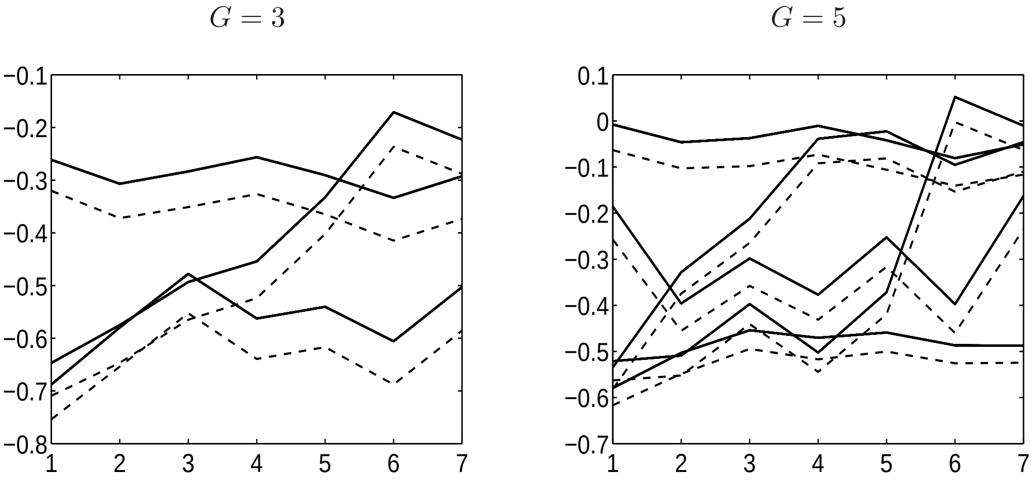
is zero, for some $(g, t) \in \{1, ..., G\} \times \{1, ..., T\}$. In this case there are no observations to compute $\alpha_{gt}^{(s+1)}$ and the algorithm stops (that is, we run it with another starting value). When using Algorithm 2, we start a local search (i.e., Step 5) as soon as $n_{gt} = 0$ for some value (g, t).

S3 Additional small-sample exercise

In this section of the supplementary appendix we provide additional Monte Carlo results. We start by comparing the estimation results of Table 3 with estimates obtained using a natural alternative: the interactive fixed-effects estimator. For the simulated dataset with $G = 3$, we estimate the interactive FE estimator of Bai (2009) allowing for three factors. Even though this estimator, like GFE, is consistent as N and T tend to infinity, the results of $1,000$ Monte Carlo replications show very substantial biases: the mean of the autoregressive parameter and the coefficient of $\widetilde{x}_{it}$ are $-.356$ and $.155$, respectively. These results suggest that the more parsimonious GFE estimator may dominate interactive FE in relatively short panels.[7]

Turning next to group-specific time effects, Figure S1 shows the pointwise means of $\widehat{\alpha}_{gt}$ across $1,000$ simulations. Both when $G = 3$ and when $G = 5$, all time profiles are shifted downwards relative to the true ones by a similar amount. The overall patterns of heterogeneity are well reproduced. In fact, we checked that the group-specific means of y_{it} and x_{it} are almost unbiased (not reported).[8]

Figure S1: Monte Carlo bias of group-specific time effects



Note: Solid line shows the true values α_{gt}^0, dashed lines show the mean of $\widehat{\alpha}_{gt}$ across $1,000$ simulations with i.i.d. normal errors. x-axis shows time $t \in \{1, ..., 7\}$.

The evidence in Section 5 is based on a design with i.i.d. normal errors, which might seem too favorable given that the asymptotic behavior of the GFE estimator crucially depends on tail and dependence properties of errors. To address this concern, we report results using a different DGP, in which errors are resampled (with replacement) from the unit-specific vectors of GFE residuals. Note that, given the nature of the original data, these residuals exhibit serial correlation and are clearly not normally distributed. Tables S1 and S2

[7] Bai (2009) discusses bias reduction in interactive FE models with strictly exogenous regressors. Moon and Weidner (2010a) provide truncation-based bias reduction formulas in models with predetermined regressors. Note that, in contrast with interactive FE, the GFE estimator is automatically higher-order bias-reducing, even in the presence of lagged outcomes or general predetermined regressors.

[8] We also have computed the finite-sample variances of the group-specific time effects, and compared them with the clustered estimator (22). As in Table 4, the results show some sizable differences between the two.

report the mean and the standard deviation and coverage of the GFE estimator for θ, respectively, across $1,000$ simulations. Compared with the i.i.d. normal case, the results show slightly larger small-sample biases, and a stronger underestimation of the finite-sample variance when using the formula based on large-T approximation. At the same time, Pollard's fixed-T formula and the bootstrap yield more accurate inference.[9]

Table S1: Bias of the GFE estimator (alternative DGP)

	θ_1 (coeff. $y_{i,t-1}$)		θ_2 (coeff. $\widetilde{x}_{it}$)		$\frac{\theta_2}{1-\theta_1}$		Misclassified
	True	GFE	True	GFE	True	GFE	
$G = 3$	.407	.381	.089	.099	.151	.163	9.86%
$G = 5$	.255	.314	.079	.082	.107	.125	13.50%
$G = 10$	.277	.322	.075	.074	.104	.109	33.27%

Note: See the notes to Table 3. Unit-specific sequences of errors are drawn with replacement from the estimated GFE residuals.

As a last exercise, we check the performance of the BIC criterion (27) to estimate the number of groups. To do so, we count the number of times that BIC selects a given G, across 100 simulated datasets. The results reported in Table S3 suggest that the criterion performs reasonably well, even in cases where the true number of groups is relatively large ($G^0 = 10$).[10] In addition, we also run simulations where the number of groups G used in estimation differs from the true number G^0. Figure S2 shows that the mean and standard deviation of the GFE estimator of common parameters do not differ much when $G > G^0$ compared to when $G^0 = 3$, consistently with the discussion in Section 5, although we observe some increase in the finite-sample dispersion of the estimator as G grows.

[9] The results for group-specific time effects are similar to those shown in Figure S1 and are omitted.

[10] We also tried the alternative choice $\widehat{\sigma}^2 \frac{G(T+N-G)}{NT} \ln(NT)$ for the penalty, instead of $\widehat{\sigma}^2 \frac{GT+N+K}{NT} \ln(NT)$ in equation (27). This corresponds to a common choice of penalty in factor models (e.g., Bai and Ng, 2002). We found that, in this case, BIC selected 1 group in all 100 simulations, when the truth was $G^0 = 3$. In comparison, Table S3 shows that our more conservative choice (27) yields superior results on these data.

Table S2: Inference for the GFE estimator (alternative DGP)

Standard errors

	θ_1 (coeff. $y_{i,t-1}$)				θ_2 (coeff. $\widetilde{x}_{it}$)				$\frac{\theta_2}{1-\theta_1}$			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$G = 3$	.050	.068	.146	.118	.0104	.0129	.0179	.0162	.011	.018	.041	.028
$G = 5$	.042	.074	.137	.125	.0083	.0108	.0126	.0103	.010	.018	.041	.033
$G = 10$	.038	.050	.091	.064	.0067	.0082	.0156	.0086	.008	.011	.026	.013

Coverage (nominal level 5%)

	θ_1 (coeff. $y_{i,t-1}$)			θ_2 (coeff. $\widetilde{x}_{it}$)			$\frac{\theta_2}{1-\theta_1}$		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
$G = 3$	.637	.792	.983	.733	.835	.994	.715	.906	.911
$G = 5$	.689	.837	.875	.887	.956	.986	.685	.855	.883
$G = 10$	.701	.862	.935	.859	.929	.989	.821	.934	.986

Note: See the notes to Tables 4. (1) *is based on the large-T variance formula,* (2) *is based on Pollard (1982)'s fixed-T formula,* (3) *is based on the bootstrap, and* (4) *in the top panel shows Monte Carlo standard deviations across simulations. Unit-specific sequences of errors are drawn with replacement from the estimated GFE residuals.*

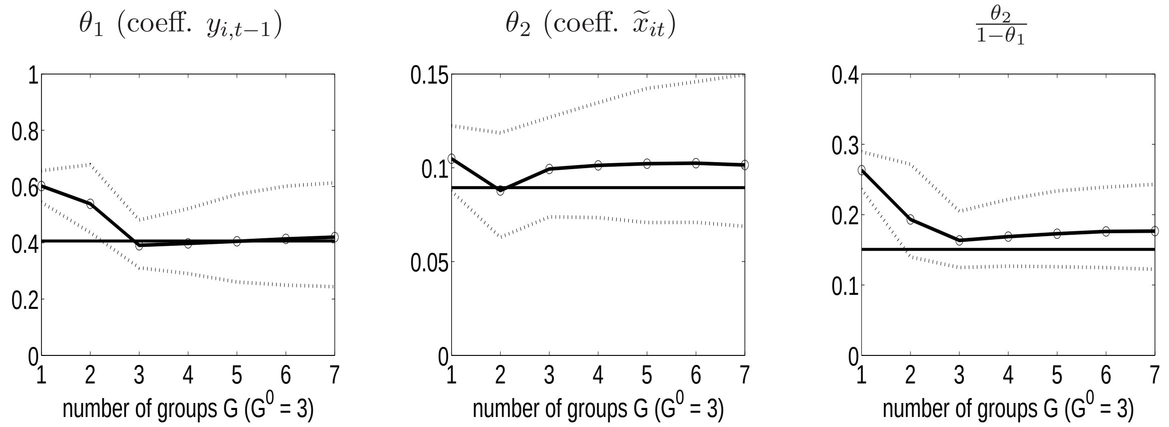
Table S3: Choice of the number of groups, BIC criterion

	$G^0 = 3$					
$G =$	1	2	3	4	5	6
$\%(\widehat{G} = G)$	0	0	98	2	0	0

	$G^0 = 10$					
$G =$	7	8	9	10	11	12
$\%(\widehat{G} = G)$	0	10	42	42	6	0

Note: See the notes to Table 3. The results show the number of times that the BIC criterion selects G groups, when the true number is $G^0 = 3$ *(upper panel) or* $G^0 = 10$ *(lower panel), respectively, out of* 100 *simulations.*

Figure S2: GFE, $G^0 = 3$, $G \neq G^0$



Note: See the notes to Table 3. The DGP has $G^0 = 3$ *groups. GFE estimates are computed using G groups, where G is reported on the x-axis. Solid thick lines and dashed lines indicate the mean and* 95% *pointwise confidence bands, respectively, across* 1,000 *simulations. The horizontal solid lines indicate true parameter values.*

References

[1] Acemoglu, D., S. Johnson, and J. Robinson (2005): "The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth," *American Economic Review*, 95, 546–79.

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

[3] Arcones, M. A., and E. Giné (1992): "On the Bootstrap of M-Estimators and Other Statistical Functionals," in *Exploring the limits of bootstrap*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., p. 1347. Wiley, New York, 1992.

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

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

[6] Giné, E., and J. Zinn (1990): "Bootstrapping General Empirical Measures," *Annals of Probability*, 18, 851–869.

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

[8] Moon, H., and M. Weidner (2010a): "Dynamic Linear Panel Regression Models with Interactive Fixed Effects," unpublished manuscript.

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

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

[11] Powell, J. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155.

[12] Silverman, B. W. (1986): *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

APPENDIX

A Proof of Proposition S1

We start with a lemma.

Lemma S1 *We have:*

$$\frac{\partial}{\partial \theta'}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\big|x_i=x\right] = \sum_{h\neq g}\left(\int_{S_{gh}} f(y|x)dy\right)\frac{(\alpha_h-\alpha_g)'}{\|\alpha_h-\alpha_g\|}x \tag{S19}$$

$$\frac{\partial}{\partial \alpha_g'}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\big|x_i=x\right] = \sum_{h\neq g}\left(\int_{S_{gh}} \frac{(y-x\theta-\alpha_g)'}{\|\alpha_h-\alpha_g\|}f(y|x)dy\right) \tag{S20}$$

$$\frac{\partial}{\partial \alpha_{\widetilde{g}}'}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\big|x_i=x\right] = -\left(\int_{S_{g\widetilde{g}}} \frac{(y-x\theta-\alpha_{\widetilde{g}})'}{\|\alpha_{\widetilde{g}}-\alpha_g\|}f(y|x)dy\right) \text{ for all } \widetilde{g}\neq g, \tag{S21}$$

where S_{gh} is given by (S5).

Proof. Let:

$$V_g=\left\{y\in\mathbb{R}^T, \|y-x\theta-\alpha_g\|^2\leq \|y-x\theta-\alpha_{\widetilde{g}}\|^2 \text{ for all } \widetilde{g}\neq g\right\}.$$

Note that:

$$\|y-x\theta-\alpha_g\|^2-\|y-x\theta-\alpha_{\widetilde{g}}\|^2=2\left(\alpha_{\widetilde{g}}-\alpha_g\right)'\left(y-x\theta-\frac{\alpha_g+\alpha_{\widetilde{g}}}{2}\right).$$

It thus follows that V_g is the intersection of $(G-1)$ half-spaces in $\mathbb{R}^T$.

We have:

$$\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\big|x_i=x\right]=\int_{V_g} f(y|x)dy.$$

Hence, using differential calculus as in Pollard (1982) we have, for all $k\in\{1,...,K\}$:

$$\frac{\partial}{\partial \theta_k}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\big|x_i=x\right]=\int_{\partial V_g} f(y|x)\nu_g(y;\theta_k)dy,$$

where ∂V_g is the frontier of V_g, and where $\nu_g(y;\theta_k)$ is the *velocity* associated to a marginal change in θ_k.

To compute $\nu_g(y;\theta_k)$, we start by noting that ∂V_g is the union of $(G-1)$ hypersurfaces:

$$\partial V_g=\bigcup_{h\neq g} S_{gh},$$

where S_{gh} is given by (S5).

As a result, we have the following identity:

$$\int_{\partial V_g} f(y|x)\nu_g(y;\theta_k)dy=\sum_{h\neq g}\int_{S_{gh}} f(y|x)\nu_g(y;\theta_k)dy.$$

Let us now define, for a given (small) $\xi\in\mathbb{R}$:

$$\theta^*=\theta+\xi e_k,$$

where e_k is a $K\times 1$ vector whose elements are all zero except a one in the kth row.

Finally, let:

$$S_{gh}^{*} = \Big\{ y \in \mathbb{R}^{T}, \|y - x\theta^{*} - \alpha_g\|^2 = \|y - x\theta^{*} - \alpha_h\|^2, \text{ and} \\ \|y - x\theta^{*} - \alpha_g\|^2 \leq \|y - x\theta^{*} - \alpha_{\widetilde{h}}\|^2 \text{ for all } \widetilde{h} \neq (g,h) \Big\}.$$

To any given $y \in S_{gh}$ we associate the point $y^{*} \in S_{gh}^{*}$ such that $y^{*} - y$ is orthogonal to the hypersurface S_{gh}. Then the velocity is defined by:

$$\nu_g(y;\theta_k) = \lim_{\xi \to 0} \frac{(y^{*} - y)' \overrightarrow{n}}{\xi},$$

where $\overrightarrow{n}$ is the normal vector to S_{gh} that points outside of V_g.

In the present case we have:

$$\overrightarrow{n} = \frac{\alpha_h - \alpha_g}{\|\alpha_h - \alpha_g\|}.$$

Moreover, y^{*} satisfies:

$$y^{*} = y + \lambda\left(\alpha_h - \alpha_g\right), \tag{S22}$$

where λ is such that:

$$(\alpha_h - \alpha_g)' \left(y^{*} - x\theta^{*} - \frac{\alpha_g + \alpha_h}{2} \right) = 0.$$

That is:

$$(\alpha_h - \alpha_g)' \left(y - x\theta - \frac{\alpha_g + \alpha_h}{2} + \lambda\left(\alpha_h - \alpha_g\right) - \xi x e_k \right) = 0,$$

from which we get:

$$\lambda = \xi \frac{(\alpha_h - \alpha_g)' x e_k}{\|\alpha_h - \alpha_g\|^2}.$$

It thus follows that:

$$\begin{aligned} \nu_g(y;\theta_k) &= \lim_{\xi \to 0} \frac{\lambda\left(\alpha_h - \alpha_g\right)' \left(\frac{\alpha_h - \alpha_g}{\|\alpha_h - \alpha_g\|}\right)}{\xi} \\ &= \frac{(\alpha_h - \alpha_g)' x e_k}{\|\alpha_h - \alpha_g\|}. \end{aligned}$$

Combining, we get:

$$\int_{\partial V_g} f(y|x)\nu_g(y;\theta_k)dy = \sum_{h \neq g} \left(\int_{S_{gh}} f(y|x)dy \right) \frac{(\alpha_h - \alpha_g)'}{\|\alpha_h - \alpha_g\|} x e_k,$$

and hence:

$$\frac{\partial}{\partial \theta'} \mathbb{E}\left[\mathbf{1}\left\{ \widehat{g}_i\left(\theta, \alpha\right) = g \right\} | x_i = x \right] = \sum_{h \neq g} \left(\int_{S_{gh}} f(y|x)dy \right) \frac{(\alpha_h - \alpha_g)'}{\|\alpha_h - \alpha_g\|} x.$$

This shows (S19).

To show (S20) and (S21) we proceed similarly. The only difference is the characterization of the velocity. We start by computing $\nu_g(y;\alpha_{gt})$ for $y \in S_{gh}$. To do this we define λ as in (S22), but now y^{*} solves:

$$(\alpha_h - \alpha_g^{*})' \left(y^{*} - x\theta - \frac{\alpha_g^{*} + \alpha_h}{2} \right) = 0,$$

where

$$\alpha_g^* = \alpha_g + \xi e_t,$$

and where with a slight abuse of notation e_t now denotes a $T \times 1$ vector whose elements are all zero except a one in the tth row.

That is:

$$(\alpha_h - \alpha_g - \xi e_t)' \left(y - x\theta - \frac{\alpha_g + \alpha_h}{2} + \lambda (\alpha_h - \alpha_g) - \frac{\xi}{2} e_t \right) = 0,$$

so:

$$\lambda = \xi \frac{(y - x\theta - \alpha_g)' e_t}{\|\alpha_h - \alpha_g\|^2} + o(\xi).$$

It thus follows that:

$$\begin{aligned} \nu_g(y; \alpha_{gt}) &= \lim_{\xi \to 0} \frac{\lambda (\alpha_h - \alpha_g)' \left(\frac{\alpha_h - \alpha_g}{\|\alpha_h - \alpha_g\|} \right)}{\xi} \\ &= \frac{(y - x\theta - \alpha_g)' e_t}{\|\alpha_h - \alpha_g\|}. \end{aligned}$$

Combining the results yields (S20).

Lastly, we compute $\nu_g(y; \alpha_{\widetilde{g}t})$ for $y \in S_{gh}$, for all $\widetilde{g} \neq g$ and all t. There are two cases:

- If $\widetilde{g} \neq h$ then $\lambda = 0$ so $\nu_g(y; \alpha_{\widetilde{g}t}) = 0$.
- If instead $\widetilde{g} = h$ then y^* solves:

$$(\alpha_h^* - \alpha_g)' \left(y^* - x\theta - \frac{\alpha_g + \alpha_h^*}{2} \right) = 0,$$

where:

$$\alpha_h^* = \alpha_h + \xi e_t.$$

That is:

$$(\alpha_h - \alpha_g + \xi e_t)' \left(y - x\theta - \frac{\alpha_g + \alpha_h}{2} + \lambda (\alpha_h - \alpha_g) - \frac{\xi}{2} e_t \right) = 0,$$

so:

$$\lambda = -\xi \frac{(y - x\theta - \alpha_h)' e_t}{\|\alpha_h - \alpha_g\|^2} + o(\xi).$$

Following the above steps yields (S21).

■

We then have the following result.

Lemma S2

$$\left. \frac{\partial}{\partial \alpha'_g} \right|_{(\overline{\theta}, \overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{ \widehat{g}_i(\theta, \alpha) = g \right\} \left(y_i - x_i \overline{\theta} - \overline{\alpha}_g \right) | x_i = x \right] = \sum_{h \neq g} \left(\int_{\overline{S}_{gh}} \frac{(y - x\overline{\theta} - \overline{\alpha}_g)(y - x\overline{\theta} - \overline{\alpha}_g)'}{\|\overline{\alpha}_h - \overline{\alpha}_g\|} f(y|x) dy \right), \tag{S23}$$

and, for all $\widetilde{g} \neq g$*:*

$$\left. \frac{\partial}{\partial \alpha'_{\widetilde{g}}} \right|_{(\overline{\theta}, \overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{ \widehat{g}_i(\theta, \alpha) = g \right\} \left(y_i - x_i \overline{\theta} - \overline{\alpha}_g \right) | x_i = x \right] = -\left(\int_{\overline{S}_{g\widetilde{g}}} \frac{(y - x\overline{\theta} - \overline{\alpha}_g)(y - x\overline{\theta} - \overline{\alpha}_{\widetilde{g}})'}{\|\overline{\alpha}_{\widetilde{g}} - \overline{\alpha}_g\|} f(y|x) dy \right). \tag{S24}$$

Proof. The lemma is a simple consequence of Lemma S1, so its proof is omitted. ■

Lastly we prove Proposition S1. We have:

$$\begin{aligned}
\Gamma_{\theta\theta} &= -\frac{\partial}{\partial\theta'}\Big|_{(\overline{\theta},\overline{\alpha})}\mathbb{E}\left[x_i'\left(y_i - x_i\theta - \alpha_{\widehat{g}_i(\theta,\alpha)}\right)\right] \\
&= \mathbb{E}\left[x_i'x_i\right] + \sum_{g=1}^{G}\mathbb{E}\left[x_i'\overline{\alpha}_g\frac{\partial}{\partial\theta'}\Big|_{(\overline{\theta},\overline{\alpha})}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}|x_i\right]\right] \\
&= \mathbb{E}\left[x_i'x_i\right] + \sum_{g=1}^{G}\mathbb{E}\left[x_i'\overline{\alpha}_g\left(\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\left(\overline{\alpha}_h-\overline{\alpha}_g\right)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}x_i\right)\right],
\end{aligned}$$

where we have used (S19). We also note that, with probability one:

$$\sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\overline{\alpha}_g\overline{\alpha}_g'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|} = \sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\overline{\alpha}_h\overline{\alpha}_h'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|},$$

since $\overline{S}_{gh}=\overline{S}_{hg}$ for all (g,h). Likewise:

$$\sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\overline{\alpha}_g\overline{\alpha}_h'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|} = \sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\overline{\alpha}_h\overline{\alpha}_g'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}.$$

Hence:

$$\begin{aligned}
\sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\overline{\alpha}_g\frac{\left(\overline{\alpha}_h-\overline{\alpha}_g\right)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|} &= \sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\overline{\alpha}_g\overline{\alpha}_h'-\overline{\alpha}_g\overline{\alpha}_g'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|} \\
&= \sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\frac{1}{2}\overline{\alpha}_g\overline{\alpha}_h'+\frac{1}{2}\overline{\alpha}_h\overline{\alpha}_g'-\frac{1}{2}\overline{\alpha}_g\overline{\alpha}_g'-\frac{1}{2}\overline{\alpha}_h\overline{\alpha}_h'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|} \\
&= -\frac{1}{2}\sum_{g=1}^{G}\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}f(y|x_i)dy\right)\frac{\left(\overline{\alpha}_h-\overline{\alpha}_g\right)\left(\overline{\alpha}_h-\overline{\alpha}_g\right)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}.
\end{aligned}$$

This shows (S6).

Next, for given $g\in\{1,...,G\}$:

$$\begin{aligned}
\Gamma_{\theta g} &= -\frac{\partial}{\partial\alpha_g'}\Big|_{(\overline{\theta},\overline{\alpha})}\mathbb{E}\left[x_i'\left(y_i - x_i\theta - \alpha_{\widehat{g}_i(\theta,\alpha)}\right)\right] \\
&= \mathbb{E}\left[x_i'\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\right] + \mathbb{E}\left[x_i'\overline{\alpha}_g\frac{\partial}{\partial\alpha_g'}\Big|_{(\overline{\theta},\overline{\alpha})}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}|x_i\right]\right] \\
&\quad + \sum_{\widetilde{g}\neq g}\mathbb{E}\left[x_i'\overline{\alpha}_{\widetilde{g}}\frac{\partial}{\partial\alpha_g'}\Big|_{(\overline{\theta},\overline{\alpha})}\mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=\widetilde{g}\right\}|x_i\right]\right] \\
&= \mathbb{E}\left[x_i'\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\right] + \mathbb{E}\left[x_i'\overline{\alpha}_g\left(\sum_{h\neq g}\left(\int_{\overline{S}_{gh}}\frac{\left(y-x_i\overline{\theta}-\overline{\alpha}_g\right)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}f(y|x_i)dy\right)\right)\right] \\
&\quad - \sum_{\widetilde{g}\neq g}\mathbb{E}\left[x_i'\overline{\alpha}_{\widetilde{g}}\left(\int_{\overline{S}_{g\widetilde{g}}}\frac{\left(y-x_i\overline{\theta}-\overline{\alpha}_g\right)'}{\|\overline{\alpha}_{\widetilde{g}}-\overline{\alpha}_g\|}f(y|x_i)dy\right)\right],
\end{aligned}$$

where we have used (S20) and (S21).

We then have:

$$\begin{aligned}
\Gamma_{gg} &= -\frac{\partial}{\partial \alpha_g'}\Big|_{(\overline{\theta},\overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\left(y_i - x_i\theta-\alpha_g\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\right] I_T - \mathbb{E}\left[\frac{\partial}{\partial \alpha_g'}\Big|_{(\overline{\theta},\overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\left(y - x_i\overline{\theta}-\overline{\alpha}_g\right)|x_i\right]\right] \\
&= \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\overline{\theta},\overline{\alpha}\right)=g\right\}\right] I_T - \mathbb{E}\left[\sum_{h\neq g}\left(\int_{\overline{S}_{gh}} \frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)(y-x_i\overline{\theta}-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|} f(y|x_i)dy\right)\right],
\end{aligned}$$

where we have used (S23).

Lastly we have, for $\widetilde{g}\neq g$:

$$\begin{aligned}
\Gamma_{g\widetilde{g}} &= -\frac{\partial}{\partial \alpha_{\widetilde{g}}'}\Big|_{(\overline{\theta},\overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\left(y_i - x_i\theta-\alpha_g\right)\right] \\
&= -\mathbb{E}\left[\frac{\partial}{\partial \alpha_{\widetilde{g}}'}\Big|_{(\overline{\theta},\overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i\left(\theta,\alpha\right)=g\right\}\left(y - x_i\overline{\theta}-\overline{\alpha}_g\right)|x_i\right]\right] \\
&= \mathbb{E}\left[\left(\int_{\overline{S}_{g\widetilde{g}}} \frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)(y-x_i\overline{\theta}-\overline{\alpha}_{\widetilde{g}})'}{\|\overline{\alpha}_{\widetilde{g}}-\overline{\alpha}_g\|} f(y|x_i)dy\right)\right],
\end{aligned}$$

where we have used (S24).

This ends the proof of Proposition S1.