Discrete Heterogeneity Patterns in Panel Data*

Stéphane Bonhomme  Elena Manresa
CEMFI, Madrid   CEMFI, Madrid

November 2011

PRELIMINARY AND INCOMPLETE

Abstract

We propose a panel data estimator that leaves the relationship between observables and unobservables unrestricted, while allowing for flexible time-varying patterns of heterogeneity. Our approach restricts the support of unobserved heterogeneity, effectively assuming that individual units belong to a small number of groups. The “grouped fixed-effects” estimator that this paper introduces is shown to be higher-order unbiased as $N$ and $T$ tend to infinity, under conditions that we characterize. As a result, inference is not affected by the fact that group membership is estimated. We apply our approach to study the link between income and democracy, while allowing for permanent and time-varying country-specific heterogeneity. The results differ significantly from approaches that assume that country heterogeneity does not vary over time.

JEL codes: C23.

Keywords: Discrete heterogeneity, panel data, fixed effects.

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, some authors have advocated the use of flexible empirical approaches to model unobserved heterogeneous features (e.g., Browning and Carro, 2005). In practice, however, there exists a trade-off between specifying rich patterns of heterogeneity, and building parsimonious specifications that are well adapted to the data at hand.

*We thank Dante Amengual, Manuel Arellano, David Dorn, Ivan Fernandez-Val, Han Hong, Bo Honoré, Andrea Ichino, Jacques Mairesse, Enrique Sentana, and seminar participants at CAM, CEMFI, EUI, Toulouse School of Economics, and CREST for useful comments. All remaining errors are our own.
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 the unobserved effects and covariates. When one is interested in measuring the effect of one particular covariate, this means that general fixed-effects endogeneity is taken care of in estimation.

Standard fixed-effects approaches are neither very parsimonious nor very flexible, however. As they allow for as many parameters as individual units, they are subject to an “incidental parameter” bias that may be substantial in finite samples (Nickel, 1981, Hahn and Newey, 2004). In addition, while standard time-invariant FE approaches provide a flexible modelling of cross-sectional heterogeneity, they are very restrictive in the time-series dimension.

A different modelling approach is to adopt a “random-effects” perspective, whereby the conditional distribution of individual effects given covariates is parametrically or semi-parametrically specified. See for example Alvarez et al. (2010) and Cunha et al. (2010) for recent applications to model time-invariant and time-varying unobserved heterogeneity, respectively. In contrast with FE, the random-effects approach restricts the nature of covariates endogeneity, typically via functional form or conditional independence assumptions.

In this paper we consider a third approach to the modelling of unobserved heterogeneity, which leaves its relationship with covariates unrestricted while allowing for flexible time-varying patterns. The main idea is to restrict the support of heterogeneity. We shall focus on the following linear model:

\[ y_{it} = x_{it}' \theta + \alpha_{it} + v_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \]  

where \( \alpha_{it} \) are unobserved unit-specific effects that are potentially correlated with covariates \( x_{it} \), while \( v_{it} \) and \( x_{it} \) are uncorrelated. In this model, our approach restricts the number of values that \( \alpha_{it} \) may take, leaving the relationship with \( x_{it} \) unrestricted.

Note that an additive FE approach would specify \( \alpha_{it} = \eta_i + \delta_t \) in model (1). In contrast, we shall consider restricting the number of distinct unit-specific sequences of unobserved effects \( \alpha_{i1}, \ldots, \alpha_{iT} \). Specifically, we will assume that individual units may be grouped into \( G \leq N \) homogeneous groups, so that:

\[ \alpha_{it} = a_{g,t}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \]  

where \( g_i \in \{1, \ldots, G\} \) denotes group membership, and \( (a_{g1}, \ldots, a_{gT}) \) are \( G \) group-specific sequences of time effects.

Importantly, no other restrictions are imposed on the unobserved effects \( \alpha_{it} \) apart from the group structure. In particular, the group membership variables \( g_i \) are left fully unrestricted. This is conceptually similar to fixed-effects approaches, in that observed and unobserved covariates are allowed to be unrestrictedly related. In comparison with the standard additive FE approach, however, model (1)-(2) restricts the amount of cross-sectional heterogeneity, while allowing for heterogeneous evolution across
groups over time. We shall call this approach “grouped fixed-effects” (GFE). The aim of this paper is to provide an estimator for the parameters of this model, and to study some of its properties. In addition, we shall provide a method to choose the number of groups $G$.

The grouped fixed-effects model (1)-(2) may be extended in various directions. One possibility is to impose some structure on the time-series evolution of the group-specific trends $a_{gt}$. Another possibility is to allow for unit-specific effects in addition to the time-varying group effects, yielding: $a_{it} = \eta_i + a_{gt}$. A third extension concerns the possibility to allow for prior information about the groups in estimation. We show that all these various modifications of the original model may easily be accommodated within our framework.

We apply our approach to study the link between income and democracy on a panel of countries. There is considerable evidence of positive association in the cross-section (Barro, 1999). Recently, Acemoglu et al. (2008) have found that this correlation disappears using panel data, when controlling for country and time-effects. We shall revisit this evidence by allowing for time-varying group effects. In this context, group-specific patterns of heterogeneity allow for time-varying determinants of political institutions that affect several countries in a similar way. As an example, the group structure allows for a “democratic wave” affecting a group of countries at one moment in time.

The grouped fixed-effects approach is related to, but different from, finite mixture modelling. Finite mixtures are widely used in economic and statistical applications.\footnote{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.} Finite mixture models rely on assumptions that restrict the relationship between unobserved heterogeneity and observed covariates.\footnote{Geweke and Keane (2007) and Norets (2010) are recent examples of flexible modelling strategies.} In contrast, our approach leaves that relationship unspecified and allows for unrestricted group probabilities. In this perspective, we argue that the grouped fixed-effects approach represents a point of contact between finite mixture modelling and fixed-effects methods.

We propose to estimate the parameters of model (1)-(2) using a restricted least squares approach. The grouped fixed-effects estimator minimizes the sum of squared residuals, subject to a constraint that restricts the number of distinct sequences of individual effects. The nature of the constraint makes computation of the global minimum a challenging task. We exploit connections with the literature on data clustering, and in particular the kmeans algorithm (Lloyd, 1957), to develop a simple and efficient iterative computation algorithm. The experiments that we report suggest that the algorithm works very well for small to moderate values of $G$.

As is commonly the case in panel data models with unobserved heterogeneity, the grouped fixed-effects estimator suffers from an incidental parameter problem. As a result, it is inconsistent when $T$ is fixed, as $N$ tends to infinity. As both dimensions of the panel increase simultaneously, however,
we show that its small-\(T\) bias vanishes rapidly, at a rate that depends on the tails of the error distribution. Hence, provided sufficiently many moments of \(v_{it}\) exist, the GFE estimator is automatically (higher-order) bias-reducing. This in turn implies that, under suitable conditions, our estimator is asymptotically equivalent to an infeasible least squares target that depends on the unknown population groups. Under these assumptions, the asymptotic distribution of GFE is thus unaffected by the fact that the group membership variables have been estimated.

Up to our knowledge, we are the first to provide this type of asymptotic characterization for models with discrete heterogeneity. The asymptotic properties of the GFE estimator contrast with available results for models with individual-specific fixed effects, where the incidental parameter bias is generally of the \(O(1/T)\) order.\(^3\) At the heart of the difference is the fact that group misclassification diminishes very fast as the number of time periods increases. Interestingly, adding non-dogmatic prior information to GFE does not affect the large-\(T\) properties of the estimator, in sharp contrast with fixed-effects models (Arellano and Bonhomme, 2009).

The application to country data on income and democracy shows evidence of time-varying country-specific heterogeneity. In addition, the effect of income strongly depends on how heterogeneity is modelled. Consistently with the findings of Acemoglu et al. (2008), we find that the income effect becomes insignificant from zero as the number of groups increases, if we impose that heterogeneity remains constant over time. This conclusion is not robust to allowing for group-specific time-varying heterogeneity, however. As an example, in our preferred specification (which corresponds to an optimally chosen number of groups) a 10% increase in income per capita is associated with a 1% increase in the Freedom House indicator. This is a small but statistically significant effect, which represents 40% of the pooled OLS estimate.

The grouping of countries that we obtain is also informative. Our estimation results clearly distinguish stable countries of the democratic type (such as countries in Western Europe or North America) from countries of the non-democratic type (China, Iran, most of the Arab world). The grouping also identifies countries in transition towards democracy, the transition occurring in the 1970-80s (Spain, Portugal or Greece, Latin American countries) or in the 1990s (South Africa, Romania). Standard additive fixed-effects approaches fail to account for these countries in transition, which amount to a substantial share of our data (roughly 30%). Overall, the empirical results that we obtain illustrate the usefulness of grouping strategies to allow for flexible patterns of unobserved heterogeneity.

**Literature and outline.** Our approach is related to the recent literature on large factor models (Stock and Watson, 2002, Bai and Ng, 2002), and panel data models with interactive fixed-effects (Bai, 2009, 2011, Moon and Weidner, 2010). Indeed, the GFE model of unobserved heterogeneity has

---

\(^3\) See for example Nickel (1981), and more recently Hahn and Newey (2004) and Arellano and Hahn (2006).

a factor-analytic structure, as:

\[ a_{g,t} = (a_{1,t}, a_{2,t}, \ldots, a_{G_t}) \times (0, 0, \ldots, 1, \ldots, 0)' \]

We take advantage of the mathematical connection with large factor models and 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 build 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.

There are important differences between our framework and factor-analytic approaches, however. One difference is interpretation: while in a factor-analytic approach estimates of the structure of heterogeneity are recovered up to an unknown rotation, the GFE approach recovers the group structure exactly. Another difference is related to estimation: for a given number of groups (or “factors”) the GFE approach is more parsimonious than factor-analytic ones, resulting in smaller asymptotic biases under correct specification. This feature of our approach may be useful in situations where the data are not informative enough to allow for fully unrestricted interactive effects.

Lastly, this paper is not the first one to explore 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 set-up, 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 individuals into several groups. They also consider a classification algorithm closely related to our iterative scheme, although they do not derive the asymptotic properties of the corresponding estimator. Our focus differs from theirs, as we aim at estimating the coefficients associated with some covariate of interest, while controlling for (discrete) unobserved heterogeneity. In addition, unlike previous work on group structures, an important motivation of this paper is to allow for patterns of heterogeneity that vary over time.\(^4\)

The outline of the paper is as follows. In Section 2 we introduce the GFE estimator, and we describe various extensions in Section 3. In Section 4 we discuss computation issues. In Section 5 we derive the asymptotic properties of the estimator as \( N \) and \( T \) tend to infinity. Section 6 considers situations where the number of groups is unknown to the researcher and must be estimated. Sections 7 and 8 apply the approach to study the relationship between income and democracy. Lastly, Section 9 concludes.

\(^4\)Another related work is Sun (2005), who considers parametric finite mixture models for panel data, and studies the properties of maximum likelihood estimation in this context.
2 The Grouped Fixed Effects estimator

In this section we start by providing a general presentation of our estimator, which applies to nonlinear models as well as to linear ones. Then, we particularize the results in the linear model with time-varying unobservables, which is the model of interest in this paper. We end the section by providing a link between our estimator and finite mixture modelling.

2.1 A general presentation

We consider a class of panel data models that relates a vector of outcomes $y_{it} = (y_{i1},...,y_{iT})'$ to a matrix of regressors $x_{it} = (x_{i1},...,x_{iT})'$, where $\dim x_{it} = K$. Here we assume for simplicity that the panel is balanced (e.g., $T_i = T$ for all $i$), although this is not essential to the derivations.

The model contains two types of parameters: a parameter vector $\theta \in \Theta$, that is common across individuals, and a collection of individual-specific parameter vectors $\alpha_i \in A$, for $i \in \{1,...,N\}$. For conciseness we will denote $\alpha = (\alpha_1,\ldots,\alpha_N)'$. The parameter spaces $\Theta$ and $A$ are subsets of $\mathbb{R}^K$ and $\mathbb{R}$, respectively. An interesting special case will be when $\alpha_i = (\alpha_{i1},\ldots,\alpha_{iT})'$ is a sequence of individual effects.

Let $m_{it}(\cdot)$ be a model-specific function that depends on the data $(y_{it}, x_{it}')$. To fix ideas, one may think of a least-square objective: $m_{it}(\theta, \alpha_i) = (y_{it} - x_{it}'\theta - \alpha_i)^2$, although the following representation allows for general M-estimators.\(^5\) The Grouped Fixed-Effects (GFE) of $(\theta, \alpha)$ is given by:

$$\left(\hat{\theta}, \hat{\alpha}\right) = \arg\min_{(\theta, \alpha) \in \Theta \times A^N} \sum_{i=1}^{N} \sum_{t=1}^{T} m_{it}(\theta, \alpha_i),$$

s.t. $\#\{\alpha_1,\ldots,\alpha_N\} \leq G,$

where $\#\{\alpha_1,\ldots,\alpha_N\}$ denotes the cardinality of the set $\{\alpha_1,\ldots,\alpha_N\}$. In words, the GFE estimator minimizes a familiar objective function, subject to the restriction that at most $G$ of the $N$ individual-specific parameter vectors $\alpha_i$ are distinct. The number $G \geq 1$ is specified by the researcher. In Section 6 we will discuss how to choose $G$ based on the data.

Two special cases are worth mentioning from the outset. If $G = 1$ then the objective is minimized subject to the restriction that all $\alpha_i$ are equal. This amounts to assuming away unobserved individual heterogeneity. At the other extreme, if $G = N$ then the constraint in (3) is not restrictive, and $\left(\hat{\theta}, \hat{\alpha}\right)$ is a standard “fixed-effects” estimator obtained by jointly minimizing the objective over common parameters and (unrestricted) individual effects.

For the purpose of computing the estimator in practice and establishing its asymptotic properties, it will be convenient to provide a slightly different characterization of the GFE estimator. To proceed,\(^5\)Other examples could be a nonlinear least squares specification of the form: $m_{it}(\theta, \alpha_i) = (y_{it} - h(x_{it}; \theta, \alpha_i))^2$, or a likelihood objective of the form: $m_{it}(\theta, \alpha_i) = -\ln f(y_{it}|x_{it}; \theta, \alpha_i)$.
notice that the constraint in (3) may equivalently be written as:

\[ \alpha_i = a_{g_i}, \quad \text{for all } i = 1, \ldots, N, \quad (4) \]

for some \( a = (a_1, \ldots, a_G)' \in \mathcal{A}^{Gq} \), and some \( (g_1, \ldots, g_N) \in \{1, \ldots, G\}^N \). This means that, under the constraint, individuals are effectively grouped into \( G \) homogeneous groups, the group variable \( g_i \) mapping individual \( i \) to one of the \( G \) groups.

Using the parameterization (4) one can define, given some parameter values \( (\theta, a) \in \Theta \times \mathcal{A}^{Gq} \), the following estimates of the group variables:

\[ \hat{g}_i(\theta, a) = \arg\min_{g \in \{1, \ldots, G\}} \sum_{t=1}^{T} m_{it}(\theta, a_g), \quad \text{for all } i = 1, \ldots, N, \quad (5) \]

taking the minimum \( g \) in case of a non-unique solution.

Finally, the GFE estimate of \( (\theta, a) \) is characterized as the minimizer of the following objective, where the group variables have been concentrated out:

\[ (\hat{\theta}, \hat{a}) = \arg\min_{(\theta, a) \in \Theta \times \mathcal{A}^{Gq}} \sum_{i=1}^{N} \sum_{t=1}^{T} m_{it}(\theta, a_{\hat{g}_i(\theta, a)}), \quad (6) \]

where \( \hat{g}_i(\theta, a) \) is given by (5).

Note that the GFE estimate of \( \alpha \) is then simply obtained as:

\[ \hat{\alpha}_i = \hat{a}_{\hat{g}_i(\hat{\theta}, \hat{a})}, \quad \text{for all } i = 1, \ldots, N. \quad (7) \]

The equations (6)-(7) represent the grouped fixed-effects estimator as a minimizer of a concentrated objective function. Note that the objective function on the right-hand side of (6) is invariant to permutations of the set \( \{a_1, \ldots, a_G\} \). This reflect the fact that the labelling of the groups \( g_i \in \{1, \ldots, G\} \) is arbitrary. Our consistency proof for \( \hat{a} \) in Section 5 will explicitly address this issue.

Lastly, the above representation of the estimator is not limited to linear models. For example, likelihood models such as static or dynamic probit or tobit models are covered. Studying the properties of GFE in nonlinear models raises some challenges, however, and we leave this issue to future work. In the rest of the paper we shall restrict ourselves to a linear model with time-varying heterogeneity, which we present next.

### 2.2 The linear model

We will consider the following linear model with varying heterogeneity:

\[ y_{it} = x_{it}'\theta + a_{g_i} + v_{it}. \quad (8) \]

In this case, \( a_g \) is a \( T \)-dimensional vector (i.e., \( q = T \)). Note that, with unrestricted group-specific time effects, a fixed-effects approach that allows for \( G = N \) is clearly infeasible here.
It is assumed that \( E(x_{it}v_{it}) = 0 \). In particular, the covariates vector \( x_{it} \) may include strictly exogenous regressors, lagged outcomes, or general predetermined regressors. In fact, this model also allows for time-invariant regressors under certain support conditions. We defer a more precise statement of the required assumptions until Section 5.6

Taking \( m_{it}(\theta, \alpha_i) = (y_{it} - x_{it}'\theta - \alpha_{it})^2 \) in (5) we obtain the estimated group variables as:

\[
\hat{g}_i(\theta, a) = \arg\min_{g \in \{1, \ldots, G\}} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta - a_{gt})^2, \tag{9}
\]

for all \( \theta \in \Theta \) and all \( GT \times 1 \) vector \( a = (a_{11}, a_{12}, \ldots, a_{1T}, a_{21}, \ldots, a_{G1}, \ldots, a_{GT})' \).

Moreover, the GFE estimator of \( (\theta, a) \) is given by:

\[
(\hat{\theta}, \hat{a}) = \arg\min_{(\theta, a) \in \Theta \times A} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta - a_{\hat{g}_i(\theta, a)\tilde{t}})^2. \tag{10}
\]

It is apparent from (10) that the grouped fixed-effects estimator minimizes a piecewise-quadratic function. To see this, note that the values of \( \hat{g}_i(\theta, a) \) for \( i = 1, \ldots, N \) define a partition of the parameter space. On each element of this partition, the GFE objective is a simple quadratic function, corresponding to the least squares objective in the regression of \( y_{it} \) on \( x_{it} \) and interactions of group and time dummies. In particular, although the objective in (10) is globally continuous, it is neither globally differentiable nor convex as soon as \( G > 1 \).

Moreover, the number of partitions of \( N \) units into \( G \) groups increases steeply with \( N \), making exhaustive search virtually impossible. As a result of its complexity, the GFE objective may have a large number of local minima. In Section 4, we will exploit a connection with the literature on data clustering in order to address this computational difficulty.

### 2.3 A finite mixture interpretation

To end this section, we note that the grouped fixed-effects estimator may be interpreted as maximizing the (pseudo) likelihood of a finite mixture model. Making the link with finite mixtures is insightful, as finite mixture modelling is widely used in economic and statistical applications. In addition, the link with finite mixtures may be useful for computation purposes, as we will argue in Section 4.

We shall conduct the discussion in the case of the linear model (8), although the equivalence applies to nonlinear models also. To state the equivalence result, let \( \sigma > 0 \) be a scaling parameter. Then, it

---

6When covariates are endogenous but instruments are available (i.e., \( E(x_{it}v_{it}) \neq 0 \) and \( E(z_{it}v_{it}) = 0 \)), one can apply the GFE approach separately to the first stage: \( x_{it} = \Pi z_{it} + b_{g_i,t} + \varepsilon_{it} \), and to the reduced-form equation (with \( \psi = \Pi'\theta \)):

\[ y_{it} = z_{it}'\psi + c_{g_i,t} + u_{it}, \]

and take \( \hat{\theta} = (\Pi\Pi')^{-1}\Pi\psi \).
is easy to see that the GFE estimator of \((\theta, a)\) satisfies:

\[
\left(\hat{\theta}, \hat{a}\right) = \arg\max_{(\theta, a) \in \Theta \times A^G T} \left[ \max_{\pi_1, \ldots, \pi_N} \sum_{i=1}^N \ln \left( \sum_{g=1}^G \pi_{ig} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{t=1}^T (y_{it} - x_{it}'\theta - a_{gt})^2 \right) \right) \right],
\]

where the maximum is taken over all probability vectors \(\pi_i\) in the unit simplex of \(R^G\). The result comes from the fact that the individual-specific \(\pi_i\) are unrestricted in (11).\(^7\) Note also that the identity (11) holds for any choice of \(\sigma\).

Identity (11) shows that the GFE estimator may be interpreted as the maximizer of the pseudo-likelihood of a mixture-of-normals model, where the mixing probabilities are individual-specific and unrestricted. 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. Finite mixture models have been widely used in economic applications. Most empirical work postulates a parametric model for the group probabilities. See Keane and Wolpin (1997) for an influential application. See also Sun (2005) for a class of model closely related to this paper, and Geweke and Keane (2007) and Norets (2010) for flexible parametric and semiparametric mixture models.

In comparison, in the grouped fixed-effects approach the group probabilities \(\pi_{gi} = \pi_g(i)\) are unrestricted functions of the individual dummies. In this perspective, the grouped fixed-effects approach may thus be viewed as a point of contact between the finite mixture approach and the fixed-effects approach.

### 3 Extensions

Here we describe various extensions of the grouped fixed-effects framework. We start by discussing a restricted version of GFE that imposes structure on the group-specific time effects. We then describe how to allow for within-group unit-specific heterogeneity. Lastly, we show how to incorporate prior information on the groups, when available, in estimation.

#### 3.1 Imposing time patterns

Note that, compared with a standard additive fixed-effects approach, model (8) restricts the amount of cross-sectional heterogeneity while allowing for unrestricted variation in the time-series dimension.

\(^7\)Specifically, given \((\theta, a)\) values the maximum is achieved at:

\[
\hat{\pi}_i(\theta, a) = \arg\max_{\pi_i} \sum_{g=1}^G \pi_{ig} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{t=1}^T (y_{it} - x_{it}'\theta - a_{gt})^2 \right),
\]

yielding:

\[
\hat{\pi}_{ig}(\theta, a) = 1 \{ \hat{g}_i(\theta, a) = g \}, \quad \text{for all } g.
\]
In practice, one may consider model specifications that also restrict the time-series variation of the unobserved effects, such as:

\[ y_{it} = x_{it}' \theta + \sum_{r=1}^{R} a_{g_{i}r}(r) \psi_{r}(t) + v_{it}, \]  

(12)

where \( \psi_{1}, \ldots, \psi_{R} \) is a family of known univariate functions, and \( a_{g_{i}r}(r) \) are unknown scalar parameters.

Model (12) imposes some structure on the group-specific trends. As an example, one may allow for linear group-specific trends and time effects only using:

\[ y_{it} = x_{it}' \theta + a^{(1)}_{g_{i}} + a^{(2)}_{g_{i}} t + \delta_{t} + v_{it}. \]

As another example, one may impose that unobserved individual effects are permanent over time:

\[ y_{it} = x_{it}' \theta + a_{g_{i}} + \delta_{t} + v_{it}. \]

(13)

The GFE estimator equally applies to the class of restricted models (12). Moreover, the computation and asymptotic analysis of the estimator in this class of models follow as an application of the treatment of model (8).

3.2 Allowing for within-group heterogeneity

Model (8) imposes that, within a group, the variation in the unobservable \( v_{it} \) is uncorrelated with the observed covariates. In particular, this modelling rules out the presence of unit-specific fixed effects. A straightforward extension of the model is to allow for time-invariant fixed effects and time-varying grouped effects in the following fashion:

\[ y_{it} = x_{it}' \theta + a_{g_{i}t} + \eta_{i} + v_{it}, \]

(14)

where \( \eta_{i}, i = 1, \ldots, N \), are \( N \) unrestricted parameters. Denoting \( \Delta y_{it} = y_{it} - y_{i,t-1} \), (14) yields the first-differenced equation:

\[ \Delta y_{it} = \Delta x_{it}' \theta + \Delta a_{g_{i}t} + \Delta v_{it}, \]

(15)

which may be estimated using GFE when \( x_{it} \) are strictly exogenous.\(^8\)

Note that, in contrast with the benchmark GFE model, model (14) has both discrete and continuous heterogeneity. As a result, within a group \( g_{i} \) all units may have different intercepts. An alternative modelling is to assume that the amount of heterogeneity within a group \( g_{i} \) has also a grouped structure, as in the following model:

\[ y_{it} = x_{it}' \theta + a_{g_{i}t} + b_{g_{i}h_{i}} + v_{it}. \]

(16)

Model (16) contains two different layers of heterogeneity. The first layer is composed of the groups \( g_{i} \in \{1, \ldots, G\} \), which have time-varying effects \( (a_{g_{i}t}) \). The second layer consists of allowing for at

---

\(^8\)When covariates are predetermined, one may use the IV counterpart of GFE outlined in footnote 6, using lagged regressors as instruments.
most \( H_g \) different intercepts \( b_{g,h} \) within a group \( g \). The subgroup membership variables are denoted as \( h_i \in \{1,\ldots,H_g\} \).

Note that model (16) is a restricted version of model (8). This is due to the fact that \((g_i, h_i)\) maps unit \( i \) to one of \( \sum_{g=1}^{G} H_g \) groups. Compared with an unrestricted GFE model with \( \sum_{g=1}^{G} H_g \) groups, the two-layer model (16) imposes linear equality constraints on the group-specific time trends.\(^9\) As a consequence, a GFE estimator for model (16) is simply obtained by imposing these linear constraints in the minimization problem (10).

### 3.3 Incorporating prior assumptions

Finally, in certain applications researchers may possess some information about the structure of unobserved heterogeneity. For example, in some cross-country applications it may seem reasonable that OECD and non-OECD countries behave similarly in some dimensions that are unobserved to the econometrician. In such a situation, one possibility would be to impose the group structure on the data by assumption, e.g. by controlling for an OECD dummy possibly interacted with time effects. Another possibility is to combine \textit{a priori} information with data information using a penalized grouped fixed-effects estimator, as we now explain.

To proceed, suppose that the \textit{a priori} information takes the form of prior probabilities on group membership, \( \pi_{ig} \) denoting the prior probability that unit \( i \) belongs to group \( g \). A penalized GFE estimator of \((\theta, a)\) is:

\[
(\hat{\theta}^{(\pi)}, \hat{a}^{(\pi)}) = \arg\min_{(\theta,a)\in\Theta \times A^G} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x'_{it}\theta - a_{g_{i}(\pi)\theta} - a_{gt}(\pi)\right)^2,
\]

where the estimated group variables are now:

\[
\hat{g}_{i}(\pi) = \arg\min_{g\in\{1,\ldots,G\}} \sum_{t=1}^{T} \left( y_{it} - x'_{it}\theta - a_{gt}(\pi)\right)^2 - C \ln \pi_{ig},
\]

and where \( C > 0 \) is a penalty term. The penalty specifies the respective weights that prior and data information have in estimation.

Computation of the penalized GFE estimator is very similar to that of the GFE estimator given by (10). In addition, the penalized and unpenalized GFE estimators will be asymptotically equivalent under the conditions given in Section 5, provided prior information be non-dogmatic.

### 4 Computation

Computation of the grouped fixed-effects estimator is particularly challenging, given the potentially very large number of local minima of the objective function. To address this issue, we shall exploit

---

\(^9\)Indeed, according to model (16) the \((g,h)\) effect at time \( t \) is: \( a_{gt} + b_{gh} \). This imposes \( \sum_{g=1}^{G} (H_g - 1)(T - 1) \) constraints on the \( \sum_{g=1}^{G} H_g T \) group-specific time effects.
connections with the literature on clustering algorithms.

To see why minimizing the GFE objective function may be interpreted as a clustering problem, it is useful to consider the simple case where \( \theta = 0 \), i.e. when there are no covariates in the model. In this case, the minimization problem in (10) has a simple geometric interpretation, as it amounts to finding a collection of “centers” \( a_1, a_2, \ldots, a_G \) in \( \mathbb{R}^T \) such that the sum of the Euclidean distances between \( y_i \) and the closest center \( a_g \) is minimal. Indeed, in this case (10) boils down to:

\[
\hat{a} = \arg\min_{a \in \mathcal{A}^G} \sum_{i=1}^{N} \left( \min_{g \in \{1, \ldots, G\}} \sum_{t=1}^{T} (y_{it} - a_{gt})^2 \right). 
\]  

(19)

Although the minimization problem in (19) is theoretically difficult,\(^{10}\) a number of efficient heuristic methods have been proposed that show good performance in various situations. The iterative computation algorithm that we propose below is based on an extension of the most popular of these heuristic methods: the well-known kmeans algorithm (Lloyd, 1957, MacQueen, 1967).

4.1 The iterative algorithm

Let us consider the minimization problem in (10). We propose to solve for \((\theta, a)\) using the following iterative scheme.

**Algorithm 1 (iterative)**

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

2. Compute for all \( i \in \{1, \ldots, N\} \):

\[
\hat{g}_i^{(s+1)} = \arg\min_{g \in \{1, \ldots, G\}} \sum_{t=1}^{T} \left( y_{it} - x_{it}'\theta^{(s)} - a_{gt}^{(s)} \right)^2. 
\]  

(20)

3. Compute:

\[
\left( \theta^{(s+1)}, a^{(s+1)} \right) = \arg\min_{(\theta, a) \in \Theta \times \mathcal{A}^{GT}} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x_{it}'\theta - a_{g_i^{(s+1)}t})^2. 
\]  

(21)

4. Iterate on \( s \) until convergence.

The iterative algorithm alternates between an assignment step (20), where each individual \( i \) is assigned to group \( \hat{g}_i^{(s+1)} \), and an update step (21), where \( (\theta, a) \) are updated given the group assignment. The update step consists of an OLS regression which treats interactions of group indicators \( 1 \{ g_i^{(s+1)} = g \} \) and time dummies as regressors.

\(^{10}\)It has been shown that the combinatorial problem (19) is NP hard in general (e.g., Aloise et al., 2009).
In the absence of covariates (i.e., when $\theta = 0$) iterating back and forth between (20) and (21) corresponds to the standard kmeans algorithm (Lloyd, 1957). In the assignment step (20), each data point $(y_{i1},...,y_{iT})'$ is assigned to the closest group mean in $\mathbb{R}^T$. In the update step (21), the means are adjusted to correspond to the averages of the $G$ groups.

Our iterative algorithm is thus an extension of the kmeans algorithm to models with covariates. Here we apply the algorithm to the linear model with time-varying unobserved heterogeneity (8). Similar iterative schemes will apply to more general (possibly nonlinear) models, as in the general formulation (6). Recently, Li and Ng (2011) considered a related kmeans-like iterative algorithm in a random coefficients model with grouped heterogeneity.

As in kmeans, it is easy to see that the concentrated objective function on the right-hand side of (10) is non-increasing in the number of iterations. Moreover, numerical convergence is typically very fast. However, similarly as in kmeans, the nature of the algorithm makes it sensitive to local minima. This means that, in practice, the solution of the iterative algorithm will depend on the chosen starting values.

A simple initialization method is to choose random starting values, and to select the solution that yields the lowest objective. See Maitra, Peterson and Ghosh (2011) for a comparison of various initialization methods for the kmeans algorithm. In the application we will use the following method:\footnote{Another simple initialization scheme that we have considered it to randomly select $G+r$ individuals at random, and to set $\left(\hat{\theta}^{(0)},\hat{a}^{(0)}\right)$ 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 specify a distribution for $\theta^{(0)}$. In our experiments, we observed very little difference between the two initialization methods.}

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

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

$$ a_{gt}^{(0)} = y_{igt} - x_{igt}'\theta^{(0)}, \quad \text{for all } (g,t). $$

In our experience, this method produced good results in low dimensional problems. In the data that we use in the application (where $N = 90$, $T = 7$, and there are two regressors), we found the iterative algorithm to be very reliable, at least for small values of $G$. We now briefly describe some numerical properties of the iterative algorithm on these data.

4.2 Local minima: a graphical illustration

Figure 1 plots the two components of $\theta$ for the income and democracy data that we shall use in the empirical application. The x-axis and y-axis report the coefficients of lagged log-GDP per capita and lagged democracy, respectively. Each dot represents a convergence point of the iterative algorithm, for 10,000 random choices of the starting values.\footnote{Specifically, the two components of $\theta^{(0)}$ were drawn from independent standard normal distributions.} Three out of the four panels show the GFE estimator
for the following number of groups: $G = 2, 3, 5$. The lower left panel shows the results for a penalized GFE estimator, where the prior distribution postulates that $\pi_{i1} = .9$ (respectively, $\pi_{i1} = .1$) if country $i$ belongs (resp. does not belong) to the OECD.\textsuperscript{13} The bar on the right of each graph shows the values of the objective function.

The results show that, depending on the chosen starting values, the iterative algorithm may converge to very different solutions. Moreover, the number of distinct convergence points is very large, reflecting the very large number of local minima of the objective function. The lower left panel on the figure suggests that prior information, when available, may help reducing the number of local minima. However, that graph also shows that adding prior information does not fully solve the problem.

Another interesting feature of Figure 1 is that the number of convergence points increases steeply with $G$. To provide some intuition on the reliability of the iterative algorithm as $G$ increases, Figure 2 plots the minimum objective function over the first $j$ random starting values, where $j$ is shown on the x-axis. To facilitate interpretation, we replicated the exercise 100 times, and report the mean and pointwise minimum and maximum across the 100 experiments.

Figure 2 shows a sharp contrast between $G = 5$ and $G = 8$. When $G = 5$ (left panel) there is no further improvement in the objective function after $j \approx 700$, meaning that one can achieve very good results using a relatively small number of starting values in this case. Indeed, the objective function is constant from $j = 700$ onwards (equal to 12.593), and we checked that the value did not improve even after 50,000 values had been drawn. Based on our various experiments on these data, we are confident that the algorithm picks up the global minimum of the objective for $G = 5$.

The situation changes for $G = 8$. In this case new starting values lead to improvements in the value of the objective function, up to $j = 3000$. We verified that the same situation occurs for very large numbers of starting values, such as 50,000. This suggests that the iterative algorithm is unable to find the global minimum in this case, at least for this order of magnitude in the number of starting values.

### 4.3 Alternative computation methods

Improving the computation of GFE is important in problems whose dimensions ($N, T, G$ and the number of covariates) are large. Recent advances for improving the kmeans algorithm include elaborate initialization methods (Bradley and Fayyad, 1998), or methods based on simulated annealing or genetic algorithms; see the survey by Xu and Wunsch (2005). Here we briefly mention another possibility based on Markov Chain Monte Carlo techniques, which we have found useful for exploring the shape of the objective function.

\textsuperscript{13}To set the penalty term $C$ in (18) we used $C = 2\sigma^2$, where $\sigma^2$ is approximated by the sum of squared residuals of OLS ($\sigma^2 \approx .039$). This choice is motivated by the special case of the normal linear model, where the sum of squared residuals in (18) is scaled by $-1/2\sigma^2$. 14
Figure 1: Solutions of the iterative algorithm

$G = 2$  
$G = 3$

$G = 2$ (prior)  
$G = 5$

Note: Balanced panel from Acemoglu et al. (2008), 1970-2000 ($N = 90, T = 7$). “Income” is log-GDP per capita, “democracy” is measured by the Freedom House indicator. Each dot on the figure shows a convergence point of the iterative algorithm, for 10,000 random choices of starting values. The bar on the right of each graph shows the values of the objective function.
Figure 2: Objective function in the iterative algorithm

\[ G = 5 \quad G = 8 \]

Note: See note to Figure 1. The y-axis reports the value of the minimum objective function over the first \( j \) random starting values, where \( j \) is shown on the x-axis. The thick solid line shows the average across 100 experiments, while the thin lines show the pointwise minimum and maximum.

Building on Chernozhukov and Hong (2003), the idea is to sample parameter draws from the following quasi-posterior distribution:

\[
p(\theta, a|y, x) \propto h(\theta, a) \exp \left( -\frac{1}{2\sigma^2} \sum_{t=1}^{T} \sum_{i=1}^{N} \left( y_{it} - x_{it}' \theta - a_{g(\theta, a)t} \right)^2 \right), \tag{22}
\]

where \( h(\cdot) \) is a weight function supported on \( \Theta \times A^G \), where \( \sigma > 0 \) is a scale parameter, and where \((y, x) = (y_1, ..., y_N, x_1, ..., x_N)\) denotes the data sample.

Once a sequence of draws \( (\theta^{(j)}, a^{(j)}) \) from an ergodic Markov Chain is available, the GFE estimator may be approximated by the joint mode of the sequence, provided the weight function \( h(\cdot) \propto 1 \) is diffuse.\(^{14}\) Another use of the chain may be to provide good starting values for the iterative algorithm.

This approach requires to efficiently draw from the quasi-posterior distribution (22). For this purpose, it is useful to exploit the connection with finite mixture modelling that we described at the end of Section 2. In Appendix C we describe two algorithms based on Markov Chain Monte Carlo techniques to draw from the quasi-posterior.

\(^{14}\)Here we have assumed that \( \sigma > 0 \) is fixed. An alternative would be to endow \( \sigma \) with a prior distribution, and sample draws from \((\theta, a, \sigma)\). See Appendix C for details.
5 Asymptotic properties

In this section we characterize the asymptotic properties of the grouped fixed-effects estimator as $N$ and $T$ tend to infinity simultaneously. The main result of this section is Theorem 2, which shows that the incidental parameter bias of the GFE estimator vanishes rapidly as $T$ increases under suitable conditions. We start by establishing consistency.

5.1 Consistency

We consider the following data generating process:

$$y_{it} = x_{it}'\theta^0 + a^0_{g_{it}t} + v_{it}. \quad (23)$$

Here the $^0$ superscripts refer to true parameter values. In particular, $g^0_{i} \in \{1, ..., G\}$ denote the true group membership indicators. It is assumed in this section that the true number of groups $G = G^0$ is known. We shall relax this assumption in Section 6.

We will need the following conditions to ensure that the GFE estimator is consistent to the true parameter values as $N$ and $T$ tend to infinity simultaneously.

**Assumption 1 (consistency)**

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

(b) $\mathbb{E}[v_{it}x_{it}] = 0$, and $\mathbb{E}[v_{it}\alpha^0_{it}] = 0$.

(c) $\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \frac{1}{T} \sum_{t=1}^{T} v_{it}v_{jt} \right| = O_p(1)$ as $N$ and $T$ tend to infinity.

(d) Let $\mathcal{G}_{N,G}$ be the set of partitions of $\{1, ..., N\}$ into $G$ groups, and let $\{g_i\} \in \mathcal{G}_{N,G}$. We denote the group-specific means of $x_{it}$ as $\overline{x}_{g_{i}g_{it}} = \frac{\sum_{i=1}^{N} 1\{g^0_i = g\} 1\{g_i = g\} x_{it}}{\sum_{i=1}^{N} 1\{g^0_i = g\} 1\{g_i = g\}}$. It is assumed that, as $N, T$ tend to infinity:

$$\rho_{\min} \left( \inf_{\{g_i\} \in \mathcal{G}_{N,G}} \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \overline{x}_{g^0_{it}g_{it}})(x_{it} - \overline{x}_{g^0_{it}g_{it}})' \right) \geq c + o_p(1), \quad (24)$$

where $c > 0$, and $\rho_{\min}(A)$ denotes the minimum eigenvalue of the matrix $A$.

(e) The processes $v_{it}$, $x_{it}$, and $\alpha^0_{it}$ satisfy suitable moment and dependence conditions that ensure that all their squares and cross-products satisfy a weak law of large numbers.

Assumption 1a requires the parameter spaces to be compact. Given the form of the objective function, it may be possible to relax this assumption, although this is not done in this paper.\footnote{In a related context, Pollard (1981) derives a consistency result for the kmeans algorithm without assuming compactness of the parameter space.}
Assumption 1b implies that \( x_{it} \) and \( \alpha_{it}^0 \) are weakly exogenous. Weak exogeneity of \( x_{it} \) and \( g_{it}^0 \) would be needed for consistency of OLS even if the groups \( g_{it}^0 \) were observed by the econometrician. Note that this assumption allows for lagged outcomes and general predetermined regressors.

Assumption 1c allows for some amount of serial and spatial correlation in the error terms of the model. The assumption is written in this form for simplicity. More primitive conditions may be found in the literature on large factor models (e.g., Stock and Watson, 2002, Bai and Ng, 2002).

Assumption 1d is analogous to a full rank condition in standard regression models. It requires that \( x_{it} \) show sufficient variation over time and across individuals.\(^{16}\) The condition will be satisfied if, for all \( g \in \{1,...,G\} \), the conditional distribution of \( (x_{i1},...,x_{iT}) \) given \( g_{it}^0 = g \) has strictly more than \( G \) points of support. Note that this may hold even when \( x_{it} \) is discrete. As an example, if \( x_{it} \) follows a non-degenerate Bernouilli distribution, i.i.d in both dimensions, then \( (x_{i1},...,x_{iT}) \) has \( 2^T \) points of support, which may well be larger than \( G + 1 \). Note also that Assumption 1d allows for time-invariant regressors, provided their support be rich enough.

Lastly, Assumption 1e is required for technical reasons. Conditions on the various processes that guarantee the validity of the law of large numbers may be found in the literature.

We have the following result.

**Theorem 1 (consistency)** Let Assumption 1a-1e hold. Then, as \( N \) and \( T \) tend to infinity: \( \hat{\theta} \xrightarrow{p} \theta^0 \), and \( \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\hat{\alpha}_{it} - \alpha_{it}^0)^2 \xrightarrow{p} 0. \)

**Proof.** See Appendix A. \( \blacksquare \)

To show that the estimator is consistent, standard methods (e.g., Newey and McFadden, 1994) do not apply. This is because the dimension of \( \alpha \) diverges as \( N \) and \( T \) tend to infinity. The proof detailed in the appendix builds on an insight by Bai (1994, 2009). The idea is to consider an auxiliary objective function, whose minimum is consistent for \( \theta^0 \). The proof then consists in showing that the difference between the GFE objective function and the auxiliary one becomes uniformly small asymptotically.

Note that the proof of Theorem 1 does not require \( \alpha_{it}^0,\ldots,\alpha_{iT}^0 \) to take exactly \( G \) distinct values. When this happens and the \( G \) groups in the population are well separated, however, it is possible to derive a consistency result for the \( G \) group-specific trends \( a_g = (a_{g1},...,a_{gT})' \). For this we need the following assumptions.

**Assumption 2 (consistency of \( \hat{\alpha} \))**

(a) For all \( g \in \{1,...,G\} \):

\[ \text{plim } \frac{1}{N} \sum_{i=1}^{N} 1\{g_{it}^0 = g\} > 0. \]

\(^{16}\)Interestingly, Assumption 1d is reminiscent of Assumption A in Bai (2009).
(b) For all \((g, \tilde{g}) \in \{1, \ldots, G\}^2\) such that \(g \neq \tilde{g}\):

\[
\text{plim}_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} (a^0_{gt} - a^0_{\tilde{g}t})^2 > 0.
\]

Assumption 2 requires \(\alpha^0_1, \ldots, \alpha^0_N\) to take exactly \(G\) distinct values in \(\mathbb{R}^T\). Assumption 2a restricts the type probabilities to be positive in the limit, while Assumption 2b requires the points of support to be distinct from each other. The next corollary then shows that \(\hat{a}\) is consistent for \(a^0\) in a well-defined sense.

**Corollary 1** (consistency of \(\hat{a}\)) Let Assumptions 1a-e and 2a-b hold. Then:

\[
\max \left\{ \max_{g \in \{1, \ldots, G\}} \left( \min_{\tilde{g} \in \{1, \ldots, G\}} \frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{gt} - a^0_{gt})^2 \right), \max_{\tilde{g} \in \{1, \ldots, G\}} \left( \min_{g \in \{1, \ldots, G\}} \frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{\tilde{g}t} - a^0_{\tilde{g}t})^2 \right) \right\} \overset{p}{\to} 0.
\]

**Proof.** See Appendix A. ■

Corollary 1 says that the Hausdorff distance between \(\hat{a}\) and \(a^0\) converges to zero in probability as \(N\) and \(T\) tend to infinity. This specific choice of distance deals with the fact that the group labelling is arbitrary.

**Properties under fixed \(T\) as \(N\) tends to infinity.** Theorem 1 and its corollary establish consistency as \(N\) and \(T\) tend to infinity simultaneously. As \(N\) tends to infinity when \(T\) is kept fixed, however, the grouped fixed-effects estimator is generally inconsistent.

To see this, it is useful to consider the special case of model (8) without covariates \((\theta = 0)\). In this case GFE boils down to the kmeans estimator. For this model, Pollard (1981, 1982) provides primitive conditions under which \(\hat{a}\) tends in probability (relative to the Hausdorff distance) to a well-defined probability limit, and is root-\(N\) consistent and asymptotically normal for this probability limit. In Pollard’s analysis, \(T\) is fixed while \(N\) tends to infinity. Importantly, even in the case where model (8) is well specified, the probability limit of \(\hat{a}\) does not coincide with the true value \(a^0\).\(^{17}\)

This inconsistency result extends to the GFE estimator \((\hat{\theta}, \hat{a})\) in the linear model with covariates. In Appendix B we derive a system of moment restrictions that are satisfied by the probability limit of the GFE estimator, and use it to characterize its fixed-\(T\) bias. As we shall now show, however, the fixed-\(T\) bias of the estimator diminishes very fast as \(T\) increases, under suitable conditions. As a result, the GFE bias may be small in panel datasets with a relatively small number of time periods. Our empirical application will illustrate this remark.

\(^{17}\)A related point is made by Bryant and Williamson (1978), who propose a “classification maximum likelihood” (CML) approach for parameter estimation in a class of parametric mixture models. In the special case where the errors \(v_{it}\) are i.i.d normal in the linear model (8), their estimator coincides with GFE. They characterize the (fixed-\(T\)) asymptotic distribution of CML under high-level conditions, and argue that the CML estimator is inconsistent in general. See also Bryant (1991).
5.2 Asymptotic distribution

We now derive the asymptotic distribution of the GFE estimator of $\theta^0$. For this purpose, we will use the following representation of the GFE estimator of $(\theta^0, a^0)$:

$$
\left( \hat{\theta}, \hat{a} \right) = \arg\min_{(\theta, a) \in \Theta \times A} \hat{Q}(\theta, a),
$$

where:

$$
\hat{Q}(\theta, a) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \theta - a_{g_i(\theta, a)t} \right)^2.
$$

(25)

It is interesting to compare $(\hat{\theta}, \hat{a})$ with the following (infeasible) least-squares estimator:

$$
\left( \tilde{\theta}, \tilde{a} \right) = \arg\min_{(\theta, a) \in \Theta \times A} \tilde{Q}(\theta, a),
$$

where:

$$
\tilde{Q}(\theta, a) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \theta - a_{g_i} \right)^2.
$$

(27)

Note that $(\tilde{\theta}, \tilde{a})$ is the least-squares estimator in the (pooled) regression of $y_{it}$ on $x_{it}$ and the interactions of the group indicators $1\{g_i^0 = g\}$ and time dummies. The estimator is infeasible as it depends on the unknown group membership variables $g_i^0$. Under standard conditions, it can be shown that, as $N$ and $T$ tend to infinity simultaneously:

$$
\sqrt{NT} \left( \hat{\theta} - \theta^0 \right) \overset{d}{\rightarrow} N(0, V_{\theta}),
$$

(29)

with $V_{\theta}$ positive definite, and where $V_{\theta}$ may be consistently estimated using clustered formulas. We shall discuss variance estimation at the end of this section.

Theorem 2 below shows that, under conditions to be specified, the difference between $\hat{\theta}$ and $\tilde{\theta}$ is asymptotically negligible, so that $\sqrt{NT} \left( \hat{\theta} - \theta^0 \right)$ is asymptotically normal with the same variance-covariance matrix $V_{\theta}$. This asymptotic equivalence means that, in a large-$T$ perspective, the asymptotic distribution of $\hat{\theta}$ is not affected by the fact that the group variables $g_i^0$ have been estimated.

To proceed we will need the following assumptions.

**Assumption 3 (asymptotic normality)**

(a) For all $(g, \tilde{g}) \in \{1, \ldots, G\}^2$ such that $g \neq \tilde{g}$ there exists a constant $c_{g, \tilde{g}} > 0$ such that, for all $T$:

$$
\frac{1}{T} \sum_{t=1}^{T} \left( a_{g^0 t} - a_{\tilde{g}^0 t} \right)^2 \geq c_{g, \tilde{g}}.
$$

(b) There exists an even $q \geq 2$ such that $E[||v_{it}||^q] < \infty$. 

20
(c) Let \( M > 0 \) be a positive constant. Then, for all \( b_t \in \mathbb{R} \) such that \( |b_t| \leq M \) for all \( t \) and for all \( g \in \{1, \ldots, G\} \):

\[
\mathbb{E} \left[ \left( \sum_{t=1}^{T} b_t v_{it} \right) \right| g_0^i = g] \leq B_i T^{q/2}, \tag{30}
\]

where \( B_i \) is independent of \( T \), and where \( \frac{1}{N} \sum_{i=1}^{N} B_i = O_p(1) \).

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

(e) Let, for all \( (g,t) \in \{1, \ldots, G\} \times \{1, \ldots, T\} \), \( \bar{x}_{g,t} = \sum_{i=1}^{N} 1\{g_0^i = g\} x_{it} / \sum_{i=1}^{N} 1\{g_0^i = g\} \). As \( N \) and \( T \) tend to infinity:

\[
\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( x_{it} - \bar{x}_{g_0^i,t} \right) \left( x_{it} - \bar{x}_{g_0^i,t} \right)' P_\theta \Sigma_\theta,
\]

where \( \Sigma_\theta \) is a positive definite matrix.

Assumption 3a imposes a lower bound on the Euclidean distance between any two values \( a_0^g \). As the bound holds for all values of \( T \), this restriction is stronger than the one that we used to show consistency of \( \hat{a} \) (Assumption 2b).

Assumptions 3b and 3c impose conditions on the moment and dependence properties of the error process. Note that (30) holds if \( v_{it} \) are independent over time by the Marcinkiewicz-Zygmund inequality (e.g., Petrov, 1995). As shown by Doukhan and Louhichi (1999) this inequality is satisfied by a large class of dependent processes, which includes stationary ARMA processes as a special case.

Assumption 3d restricts covariates to have bounded support. This strong assumption could be relaxed, at the cost of restricting the dependence and tail properties of the \( x_{it} \) process in a similar way as Assumptions 3b and 3c restrict the properties of \( v_{it} \).

Lastly, Assumption 3e is necessary for the OLS estimator \( \tilde{\theta} \) to have a non-degenerate asymptotic variance.

We have the following result.

**Theorem 2** (asymptotic normality of \( \hat{\theta} \)) Let Assumptions 1a-1e, 2a-2b, and 3a-3e hold. Suppose in addition that \( N \) and \( T \) tend to infinity such that \( \frac{N}{T^{q/2}} \rightarrow 0 \). Lastly, suppose that (29) holds. Then we have, asymptotically:

\[
\sqrt{NT} \left( \hat{\theta} - \theta^0 \right) \xrightarrow{d} \mathcal{N}(0, V_\theta). \tag{31}
\]

**Proof.** See Appendix A. ■

The idea of the proof detailed in the appendix is as follows. We start by showing that the difference between \( \hat{Q}(\cdot) \) and \( \tilde{Q}(\cdot) \) may be uniformly bounded by the following average misclassification error:

\[
\frac{1}{N} \sum_{i=1}^{N} 1\{\hat{g}_i(\theta,a) \neq g_0^i\}. \tag{32}
\]
The misclassification error (32) is the source of the incidental parameter problem in models with discrete unobserved heterogeneity. The error comes from the fact that group membership is imprecisely estimated when $T$ is small. The second part of the proof consists in showing that, under the assumptions of Theorem 2, the average misclassification error vanishes at a suitable rate as $T$ tends to infinity.

Theorem 2 is concerned with the distribution of $\hat{\theta}$. We shall now derive the rate of convergence in quadratic mean of the group-specific trends $\hat{a}_g$. To do so we will assume that:

$$\frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{gt} - a_{gt}^0)^2 = O_p \left( \frac{1}{N} \right).$$

(33)

This is a natural condition in regression models for panel data, the intuition being that each group-specific time effect $a_{gt}^0$ is estimated with $N$ observations, hence (under suitable conditions) at a root-$N$ rate.

Then we have the following result.

**Corollary 2** (rate of convergence of $\hat{a}$) Let the assumptions of Theorem 2 hold, and suppose that (33) holds. Then, as $N$ and $T$ tend to infinity such that $\frac{N}{T^4} \to 0$ we have:

$$\frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{gt} - a_{gt}^0)^2 = O_p \left( \frac{1}{N} \right).$$

(34)

**Proof.** See Appendix A. ■

Theorem 2 and Corollary 2 rely on the fact that the difference between the GFE estimator and the infeasible least squares estimator becomes asymptotically negligible. We now provide a simple intuition for this result. To proceed, let us consider the linear model (13) with time-invariant heterogeneity. Suppose for simplicity that $G = 2$, that $\theta^0 = 0$ is known (no covariates), and that $v_{it}$ are iid normal $(0, \sigma^2)$. The model is then:

$$y_{it} = a_{g^0_i} + v_{it}, \quad g^0_i \in \{1, 2\}, \quad v_{it} \sim iid \mathcal{N}(0, \sigma^2) .$$

Let $a = (a_1, a_2)$ be some element of $\mathcal{A}^2$, taking $a_1 < a_2$ without loss of generality. It follows from (18) that the probability of misclassifying an individual who belongs to group 1 into group 2 is:

$$\Pr (\hat{g}_i(a) = 2 \mid g^0_i = 1) = \Pr \left( \sum_{t=1}^{T} (a^0_t + v_{it} - a_2) < \sum_{t=1}^{T} (a^0_t + v_{it} - a_1) \right) .$$

$$= \Pr \left( v_i > \frac{a_1 + a_2}{2} - a^0_1 \right) .$$

Now, in the neighborhood of $(a^0_1, a^0_2)$ we have $\frac{a_1 + a_2}{2} - a^0_1 \approx \frac{a_2 - a^0_1}{2} > 0$. In addition, $v_i \sim \mathcal{N} \left( 0, \frac{\sigma^2}{T} \right)$. As a result, the misclassification probability is approximately:

$$\Pr (\hat{g}_i(a) = 2 \mid g^0_i = 1) \approx 1 - \Phi \left( \sqrt{T} \left( \frac{a^0_2 - a^0_1}{2\sigma} \right) \right) .$$

(35)
The misclassification probability in (35) tends to zero at an exponential rate. This shows why, in this simple case, the incidental parameter problem due to group misclassification vanishes very rapidly as $T$ increases. Note that this result is specific of models with discrete heterogeneity. When $\alpha_i$ can take continuous values, in contrast, biases due to the incidental parameter problem are typically of the $O(1/T)$ order (e.g., Nickel, 1981, Hahn and Newey, 2004).

The normality assumption on $v_{it}$ is not necessary to get an exponential rate of convergence. For example, a similar result can be shown when $v_{it}$ are i.i.d and their distributions have exponential tails. The conditions of Theorem 2 allow for a more general class of distributions. Specifically, the statement of the theorem suggests that the tails of the error distribution (reflected in the number of finite moments of $v_{it}$) determine the amount of bias in the asymptotic distribution of $\hat{\theta}$.

The proof of Theorem 2 establishes that the average misclassification error converges to zero. In addition, (36) shows that the GFE estimator $\hat{\theta}$ is asymptotically equivalent to the infeasible least-squares estimator $\tilde{\theta}$. By these two properties, GFE satisfies an asymptotic “oracle” property. Related results have been recently proven in the literature on penalized estimation and variable selection (Fan and Li, 2001, Horowitz, Huang, and Ma, 2008). The originality in the present case is that none of the potential regressors $g_0^i$ is observed by the econometrician.\(^{18}\) Yet, panel data allow to infer group membership with probability approaching one as $N$ and $T$ tend to infinity, with no effect on the asymptotic distribution of the GFE estimator.

Lastly, note that the asymptotic equivalence between the GFE estimator and an infeasible least-squares estimator holds under very similar conditions for linear models with restricted time trends, such as models (12) and (13), respectively.

We end this section by briefly discussing two extensions.

**Group probabilities.** For interpretation purposes it may be useful to fit a parametric model (e.g., a multinomial logit model) to the estimated group probabilities. Indexing the model by a parameter vector $\xi$ one may want to compute:

$$\hat{\xi} = \arg\max_{\xi} \sum_{i=1}^{N} \left\{ \mathbb{1}\left( g_i \left( \hat{\theta}, \hat{a} \right) = g \right) \right\} \ln \left( p_g \left( x_i; \xi \right) \right),$$

where $p_g(x; \xi)$ are the parametrically specified group probabilities.

Similar arguments as in the proof of Theorem 2 imply that $\hat{\xi}$ will be asymptotically equivalent to the following infeasible ML estimator:

$$\tilde{\xi} = \arg\max_{\xi} \sum_{i=1}^{N} \left\{ \mathbb{1}\left( g_0^i = g \right) \right\} \ln \left( p_g \left( x_i; \xi \right) \right).$$

\(^{18}\)Moreover, recall from our discussion of the form of the objective function that the number partitions of $\{1, ..., N\}$ into $G$ groups is prohibitively large.
This result implies that model estimates (and their standard errors) that take the estimated group variables as dependent variables will be asymptotically valid.

**Prior assumptions.** Lastly, we consider the asymptotic properties of the penalized grouped fixed-effects estimator given by (17). We make the following assumption on prior probabilities.

**Assumption 4** (prior probabilities) The prior probabilities are non-dogmatic is the sense that, for some $\varepsilon > 0$:

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

We have the following result.

**Corollary 3** (asymptotic normality, penalized GFE) Let the assumptions of Theorem 2 hold, and let $\pi = \{\pi_{ig}\}$ be a set of prior probabilities that satisfies Assumption 4. Then we have, asymptotically:

$$
\sqrt{NT} \left( \hat{\theta}^{(\pi)} - \theta^0 \right) \xrightarrow{d} N(0, V_{\theta}).
$$

**Proof.** See Appendix A. ■

In models with continuous unobserved heterogeneity, adding prior information on the individual effects has typically a first-order effect on the bias of the estimator (Arellano and Bonhomme, 2009). In contrast, Corollary 3 shows that 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 of $\theta^0$.

### 5.3 Variance estimation

Theorem 2 shows that the grouped fixed-effects estimator $\hat{\theta}$ is asymptotically equivalent to the infeasible least squares estimator $\tilde{\theta}$. When observations are i.i.d. across individuals and $T$ is fixed, one may consider the variance estimator of Arellano (1987):

$$
\widehat{\text{AVar}} \left( \hat{\theta} \right) = \frac{\hat{V}_\theta}{NT} = \frac{\hat{\Sigma}_\theta^{-1}\hat{\Omega}_\theta\hat{\Sigma}_\theta^{-1}}{NT},
$$

where

$$
\hat{\Sigma}_\theta = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \left( x_{it} - \bar{x}_{g_i,t} \right) \left( x_{it} - \bar{x}_{g_i,t} \right)',
$$

$$
\hat{\Omega}_\theta = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \sum_{s=1}^T \hat{v}_{it} \hat{v}_{is} \left( x_{it} - \bar{x}_{g_i,t} \right) \left( x_{is} - \bar{x}_{g_i,s} \right)'.
$$

\footnote{Note that Theorems 1 and 2 allow for weak dependence in the cross-sectional dimension. If observations are not i.i.d across individuals, the clustered variance formula is invalid. In a related context, Bai and Ng (2006) considered a partial sample estimator, which uses a random sample of size $n << \min(N, T)$.}
where \( \hat{v}_{it} = y_{it} - x'_{it} \tilde{\theta} - \tilde{a}_{it} \) are the least squares residuals.

Hansen (2007) studies the properties of this estimator as \( N \) and \( T \) tend to infinity. He shows that, under suitable restrictions on the time-series properties of the data, Arellano’s variance estimator remains consistent as both dimensions get large.\(^{20}\)

Our asymptotic equivalence result suggests estimating the asymptotic variance of the GFE estimator \( \hat{\theta} \) using Arellano (1987)’s formula, replacing the unknown group variables \( g_i^0 \) by their GFE estimates \( \hat{g}_i(\hat{\theta}, \hat{a}) \) in (38)-(39). Likewise, we will use Arellano’s clustered formula to estimate the asymptotic variance of \( \hat{a}_{gt} \), for all \((g, t)\).\(^{21}\)

Lastly, it is important to emphasize that all results in this section assume that the true number of groups \( G = G^0 \) is known to the researcher. If \( G \) is misspecified, then the above inference results fail to hold in general. The next section deals with situations where \( G^0 \) is unknown.

### 6 Properties when the number of groups is unknown

In this section we consider two issues in turn. First, we discuss the properties of the grouped fixed-effects estimator when the number of groups is misspecified. Then, we propose a class of information criteria to choose \( G \) in practice.

#### 6.1 Effect of Misspecification of \( G \)

In practice the number of groups is unknown to the econometrician. Here we assume that there exists a true but unknown number of groups \( G^0 \), and we let \( G \) be the (possibly incorrect) postulated number of groups. Throughout this section we will assume that \( G \) is lower than a known upper limit \( G_{\text{max}} \) independent of the sample size.

There are two cases. When \( G < G^0 \), then the GFE estimator \( \hat{\theta} \) is generally inconsistent for \( \theta^0 \) if the unobserved effects are correlated with the observed covariates. This is a simple reflection of the omitted variable bias.

When \( G > G^0 \), in contrast, \( \hat{\theta} \) remains consistent for \( \theta^0 \) under the conditions of Theorem 1. This is because the proof of Theorem 1 is unchanged when \( G > G^0 \) (as opposed to \( G = G^0 \)). The analysis of the group-specific trends \( a_g \) is more involved, however, as the proof of Corollary 1 relies on the assumption that \( G = G^0 \). In Appendix B we provide conditions under which \( \hat{a} \) remains consistent for \( a^0 \) when \( G > G^0 \) (relative to the Hausdorff distance).

\(^{20}\)One limitation of Hansen’s results is that time effects are not allowed for in the analysis. However, he argues that, in his simulation experiments, the size distortions created by the presence of time effects decrease rapidly as \( N \) increases (Hansen, 2007, p. 609).

\(^{21}\)We acknowledge that, as of now, the large \( N,T \) validity of the clustered formula for making inference on the group-specific time effects is an open question.
In addition to showing consistency, it is interesting to characterize the rates of convergence of the components of the GFE estimator in situations where \( G \) is misspecified. Results on rates will be important for justifying the asymptotic validity of the information criteria that we shall consider at the end of this section. Let us first recall the result when \( G = G^0 \). In this case Theorem 2 implies that:

\[
\hat{\theta} - \theta^0 = O_p \left( \frac{1}{\sqrt{NT}} \right),
\]

while Corollary 2 states that:

\[
\left( \frac{1}{T} \sum_{t=1}^{T} \left( \hat{a}_{gt} - a^0_{gt} \right)^2 \right)^{1/2} = O_p \left( \frac{1}{\sqrt{N}} \right).
\]

We conjecture that (40) and (41) remain true when \( G > G^0 \). An intuition is proposed in Appendix B, where we discuss the case \( G^0 = 1, G = 2 \). Actually, the derivations in the appendix suggest that the GFE estimator for common parameters with \( G \) groups and the GFE estimator with \( G^0 \) groups are asymptotically equivalent as \( N \) and \( T \) tend to infinity at the rate specified by Theorem 2. This conjecture is also supported by the simulation experiments that we report in the appendix.\(^{22}\)

This discussion is closely related to misspecification of the number of factors in panel data models with interactive fixed-effects. Bai (2009) showed that common parameters and factor estimates have the same rate of convergence, even when the postulated number of factors is larger than the true one. Recently, Moon and Weidner (2010) have shown that common parameter estimates with \( G^0 \) (the true number of) factors and \( G > G^0 \) factors are asymptotically equivalent.

### 6.2 Information criteria for choosing \( G \)

To end this section we show how to consistently estimate the true number of groups \( G^0 \). For this we use a penalized information criterion of the form:

\[
\hat{I}(G) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \hat{\theta}^{(G)} - \hat{\alpha}^{(G)}_{it} \right)^2 + Gh_{NT},
\]

where the grouped-fixed-effects estimator with \( G \) groups \( \left( \hat{\theta}^{(G)}, \hat{\alpha}^{(G)} \right) \) is given by equation (10). Then, we set:

\[
\hat{G} = \arg\min_{G \in \{1, \ldots, G_{\text{max}}\}} \hat{I}(G).
\]

The asymptotic properties of this type of information criteria have been studied in the related context of large factor models (Bai and Ng, 2002) and interactive fixed-effects panel data models (Bai, 2009).

\(^{22}\)Note that the asymptotic equivalence result for \( \hat{\theta} \) does not apply to the group-specific trends \( \hat{a} \). In the special case \( G^0 = 1, G = 2 \), the results reported in Appendix B are consistent with trends having an \( O_p(T^{-1/2}) \) bias.
Let us start by taking $G < G^0$. As $\hat{\theta}^{(G)}$ is generally inconsistent for $\theta^0$ if heterogeneity and covariates are correlated, it is easy to see that, as $N$ and $T$ tend to infinity simultaneously:

$$\hat{I}(G) - \hat{I}(G^0) \xrightarrow{p} C > 0,$$

provided $h_{N,T}$ tends to zero asymptotically. As a result, $\hat{G} \geq G^0$ with probability approaching one.

To establish that $\hat{G}$ is consistent for $G^0$ we need the following result.

**Proposition 1** (number of groups) Suppose that the conditions of Theorem 2 are satisfied, and that (40) and (41) hold. In addition, suppose that $Nh_{NT}$ tends to infinity as $N$ and $T$ tend to infinity. Then $\hat{G} \leq G^0$ with probability approaching one.

**Proof.** See Appendix A.◼

Taken together, these results show that $\hat{G} \xrightarrow{p} G^0$ if $h_{NT} \to 0$ and $Nh_{NT} \to \infty$ as $N$ and $T$ tend to infinity. As an example, let us consider the well-known standard Akaike (1973) and BIC (Schwarz, 1978) criteria for this problem. 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 the following we shall take the upper bound $GT + N + K$, where $K = \dim \theta$. We define:

$$AIC(G) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \hat{\theta}^{(G)} - \hat{\alpha}_{it}^{(G)} \right)^2 + \frac{2 \hat{\sigma}^2}{NT} \left( GT + N + K \right),$$

$$BIC(G) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \hat{\theta}^{(G)} - \hat{\alpha}_{it}^{(G)} \right)^2 + \frac{\hat{\sigma}^2}{NT} \left( GT + N + K \right) \ln(NT),$$

where, in both expressions, $\hat{\sigma}^2$ is a low bias estimate of the variance of $v_{it}$.\(^\text{23}\) It follows from the above discussion that, in an asymptotic where $N$ and $T$ tend to infinity, BIC estimates the number of groups consistently.\(^\text{24}\) In contrast, AIC may be inconsistent.

### 7 Application: income and democracy revisited

In this section and the next we apply our approach to study the relationship between income and democracy across countries. We start by presenting the main empirical results. The next section will provide some additional evidence and robustness checks.

\(^\text{23}\)A possibility is to estimate $\hat{\theta}$ and $\hat{\alpha}$ using grouped fixed-effects with $G_{max}$ groups, and to take:

$$\hat{\sigma}^2 = \frac{1}{NT - G_{max}T - N - K} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \hat{\theta} - \tilde{\alpha}_{it}(\hat{\theta}, \hat{\alpha}) \right)^2.$$

\(^\text{24}\)This result contrast with the factor model studied in Bai and Ng (2002), where consistency of BIC depends on the relative rates of $N$ and $T$. The difference comes from the fact that, in factor models, the number of parameters (that is, $G(N + T - G)$, where $G$ is the number of factors) increases more steeply with the sample size.
7.1 The empirical framework

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) argue that it is important to control for factors that simultaneously affect both variables. Using a panel data of countries, they document that the positive effect of income on democracy disappears when including country-specific fixed-effects in the regression. Our aim here is to revisit this evidence by allowing for time-varying country heterogeneity in the analysis.\(^\text{25}\)

As in Acemoglu et al (2011) we will consider the following model:

\[
democracy_{it} = \theta_1 democracy_{i,t-1} + \theta_2 \log GDP_{pc_{it-1}} + \alpha_{it} + v_{it}.
\]

In particular, we will be interested in the short-run income effect \(\theta_2\) (controlling for lagged democracy), and in the cumulative income effect \(\theta_2/(1 - \theta_1)\). We shall compare and contrast the empirical results obtained with various specifications for the unobserved country-specific factors \(\alpha_{it}\).

As a motivation, Table 1 shows regression results controlling for time effects, country effects or both. Columns (1) to (3) use the unbalanced panel dataset of Acemoglu et al. (2008), which covers the period 1960−2000. The data are measured on five-year intervals, and the Freedom House measure of democracy is used.\(^\text{26}\) Columns (4) to (6) in the table show the results for the balanced subpanel on the period 1970−2000, which we shall use as our benchmark sample.\(^\text{27}\)

Columns (1) and (4) show that, when controlling for time effects only, there exists a positive and significant correlation between income and democracy on these data. The implied cumulative effect of income (\(^\text{25}\)) means that a 10\% increase in income per capita is associated with an increase in the Freedom House score of 2.5\%. Comparison with the other columns of the table shows that this result is sensitive to allowing for unobserved country heterogeneity. Columns (2) and (5) replicate the main result by Acemoglu et al. (2008), which shows that the income effect becomes statistically insignificant when including country and time fixed-effects in the regression. Interestingly, results are also qualitatively and quantitatively different when including country fixed effects only, as shown in columns (3) and (6), a result that was not reported by Acemoglu et al. In addition, comparing columns (4) and (5) shows that, on the balanced subsample, the standard error of the income effect increases substantially when controlling for country dummies. This reflects the fact that only 6.2\% of the variance of income comes from within-country variation (26.2\% for democracy).

A fixed-effects approach will adequately control for unobserved country heterogeneity if \(\alpha_{it} = \eta_i + \delta_t\) is additive in the country and time dimensions. Acemoglu et al (2011) interpret the country-

\(^{25}\)Recently, Benhabib et al. (2011) have argued that, by considering other types of panel data models (better suited to the nature of the democracy variable) and a longer dataset, the association may be reestablished.

\(^{26}\)Log GDP per capita, measured in 1990 US dollars, is taken from Maddison (2003).

\(^{27}\)The balanced subsample covers 90 different countries from the five continents. It contains approximately 85 per cent of the OECD members in the year 2000.
Table 1: Regression results

<table>
<thead>
<tr>
<th></th>
<th>Unbalanced panel</th>
<th>Balanced panel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Democracy_{t-1}</td>
<td>.706</td>
<td>.379</td>
</tr>
<tr>
<td></td>
<td>(.035)</td>
<td>(.041)</td>
</tr>
<tr>
<td>LogGDP_{per capita}_{t-1}</td>
<td>.072</td>
<td>.010</td>
</tr>
<tr>
<td></td>
<td>(.010)</td>
<td>(.035)</td>
</tr>
<tr>
<td>Implied cumulative effect of income</td>
<td>.246</td>
<td>.017</td>
</tr>
<tr>
<td></td>
<td>(.031)</td>
<td>(.056)</td>
</tr>
<tr>
<td>Observations</td>
<td>945</td>
<td>945</td>
</tr>
<tr>
<td>Countries</td>
<td>122</td>
<td>122</td>
</tr>
<tr>
<td>R-squared</td>
<td>.725</td>
<td>.796</td>
</tr>
<tr>
<td>Time dummies</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Country fixed effects</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

specific fixed-effects $\eta_i$ as reflecting historical, long-run factors that have shaped political and economic development. The assumption is that time-varying factors $\delta_t$ affect all countries in the same way. It is important to notice that country and time-effects are driving the Acemoglu et al. results, as we saw in Table 1. We shall now explore other specifications for the country-specific heterogeneity, allowing for determinants of political change that evolve over a shorter period of time. Relaxing the assumption that country effects are time-invariant makes intuitive sense in this application, as the baseline dataset spans more than thirty years.

Our main assumption is that time-varying determinants of political change affect groups of countries, as opposed to affecting each individual country separately. Specifically we will assume that $\alpha_{it} = a_{g_{it}}$, with $g_i \in \{1, ..., G\}$. We shall report results using various numbers of groups $G$. We have in mind the fact that “democratization waves” may affect a group of countries at a point in time, because they share important political or cultural features. In particular, this specification may capture spillover or contagion effects driving common evolution in political change in several countries at one moment in time. The decolonization era in the 1950s and 1960s, the revolutions in Eastern Europe at the end of 1990s, and the recent Arab spring in 2011 provide examples of the type of evolution that we have in mind.

### 7.2 Results on income and democracy

Figure 3 shows the point-estimates and standard errors of income and democracy coefficients for various values of the number of groups $G$.

The results for the coefficient of lagged democracy show that the grouped fixed-effects estimates for $G = 3$ or more groups are substantially lower than the one for $G = 1$, in the range of .3 to .4, compared to an OLS estimate of .67. This is consistent with the presence of unobserved heterogeneity that is positively correlated with democracy, causing an upward bias in OLS. The coefficient of lagged income remains rather stable as $G$ increases, with a slight decrease from .08 to .06.

These changes are reflected in the profile of the cumulative income effect (right panel), which sharply decreases from .25 in OLS to .10 for $G = 5$, and remains almost constant as $G$ increases further. Interestingly, although substantially smaller than in OLS, the cumulative income effect remains statistically significant even for large values of $G$.

We also report the estimated parameter values in Table 2. In addition, the first column in the table shows the value of the objective function, while the second column reports the BIC criterion (computed using $G_{max} = 15$). The table shows that the objective function decreases steadily as $G$ increases.

---

28 All estimates were computed using the iterative algorithm of Section 4, using 10,000 random starting values.

29 One concern is that as $G$ increases our asymptotic results offer a poor guide for the finite-sample performance of the GFE estimator. Our checks on simulated data suggest that this is not the case on these data, see Section 8. Another concern is that the reported estimates may not coincide with the global minimum of the GFE objective function, particularly for $G \geq 8$. Overall, this suggests interpreting the GFE results for large values of $G$ with some caution.
Figure 3: Income and democracy: GFE estimates

\[
\text{Lagged democracy } \left( \theta_1 \right) \quad \text{Lagged income } \left( \theta_2 \right) \quad \text{Cumulative income effect } \left( \frac{\theta_2}{1 - \theta_1} \right)
\]

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. Confidence intervals clustered at the country level are shown in dashed lines.

increases: by 50% when \( G = 5 \) compared to OLS, and by 75% when \( G = 13 \). Interestingly, comparison with the last row of the table shows that the objective function of grouped fixed-effects is lower than the one of the fixed-effects model as soon as \( G \geq 3 \). Together, these results suggest that there is a large amount of cross-country heterogeneity in these data, and that a substantial part of it is time-varying. The next subsection will provide additional evidence by describing how the estimated groups of countries evolve over time.

Lastly, Table 2 shows that the optimal number of groups according to the Bayesian information criterion is \( G = 10 \). This large number of groups is also consistent with country heterogeneity being very substantial. Allowing for so many groups in estimation makes interpretation of the groups problematic, as we shall see below. In addition, it is interesting to notice that the estimated parameter values do not vary substantially from \( G = 5 \) groups onwards.

7.3 Groups of countries: a description

The left column in Figure 4 plots the group-specific time effects \( \hat{a}_{gt} \) over time, for \( G = 2 \) to \( G = 6 \). In addition, the center and right columns show the group-specific means of democracy and income over time, respectively. We see that the estimated groups are clearly separated along the income and democracy dimensions. Moreover, the results strongly suggest the presence of time-varying country heterogeneity.
Table 2: Income and democracy: GFE estimates (in numbers)

<table>
<thead>
<tr>
<th>$G$</th>
<th>Objective</th>
<th>BIC</th>
<th>Demo ($\theta_1$)</th>
<th>Income ($\theta_2$)</th>
<th>Income ($\frac{\theta_2}{1-\theta_1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.722</td>
<td>.049</td>
<td>.665</td>
<td>.083</td>
<td>.257</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.049)</td>
<td>(.014)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>2</td>
<td>19.847</td>
<td>.046</td>
<td>.601</td>
<td>.061</td>
<td>.152</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.041)</td>
<td>(.011)</td>
<td>(0.023)</td>
</tr>
<tr>
<td>3</td>
<td>16.599</td>
<td>.042</td>
<td>.407</td>
<td>.080</td>
<td>.151</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.052)</td>
<td>(.011)</td>
<td>(0.014)</td>
</tr>
<tr>
<td>4</td>
<td>14.319</td>
<td>.040</td>
<td>.302</td>
<td>.082</td>
<td>.118</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.055)</td>
<td>(.010)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>5</td>
<td>12.593</td>
<td>.038</td>
<td>.255</td>
<td>.079</td>
<td>.107</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.051)</td>
<td>(.010)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>6</td>
<td>11.131</td>
<td>.036</td>
<td>.465</td>
<td>.064</td>
<td>.119</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.043)</td>
<td>(.008)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>7</td>
<td>10.059</td>
<td>.036</td>
<td>.403</td>
<td>.065</td>
<td>.108</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.044)</td>
<td>(.009)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>8</td>
<td>9.257</td>
<td>.035</td>
<td>.310</td>
<td>.070</td>
<td>.102</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.046)</td>
<td>(.008)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>9</td>
<td>8.427</td>
<td>.035</td>
<td>.325</td>
<td>.072</td>
<td>.106</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.041)</td>
<td>(.008)</td>
<td>(0.011)</td>
</tr>
<tr>
<td>10*</td>
<td>7.766</td>
<td>.035</td>
<td>.353</td>
<td>.068</td>
<td>.105</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.044)</td>
<td>(.008)</td>
<td>(0.0107)</td>
</tr>
<tr>
<td>11</td>
<td>7.279</td>
<td>.035</td>
<td>.269</td>
<td>.072</td>
<td>.099</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.039)</td>
<td>(.008)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>12</td>
<td>6.946</td>
<td>.036</td>
<td>.327</td>
<td>.058</td>
<td>.086</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.047)</td>
<td>(.009)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>13</td>
<td>6.478</td>
<td>.036</td>
<td>.252</td>
<td>.068</td>
<td>.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.038)</td>
<td>(.009)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>14</td>
<td>6.163</td>
<td>.036</td>
<td>.275</td>
<td>.058</td>
<td>.080</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.038)</td>
<td>(.008)</td>
<td>(0.010)</td>
</tr>
<tr>
<td>15</td>
<td>5.927</td>
<td>.037</td>
<td>.430</td>
<td>.054</td>
<td>.094</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.040)</td>
<td>(.008)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>FE</td>
<td>17.517</td>
<td>−</td>
<td>.284</td>
<td>−.031</td>
<td>−.044</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(.058)</td>
<td>(.049)</td>
<td>(0.069)</td>
</tr>
</tbody>
</table>

Note: See the notes to Figure 3. The table reports the value of the objective function, the Bayesian information criterion, and coefficient estimates with their standard errors for the GFE estimates with various values for the number of groups $G$. The parameter $\hat{\sigma}^2$ in BIC was computed using $G_{\text{max}} = 15$. The last row in the table shows the same figures for the fixed-effects model.
As an example, let us consider the case $G = 3$ (second row in Figure 4). While two of the groups experience parallel evolution in democracy (slightly increasing over the period), the third group experiences a remarkable increase in democracy from 1970 to 2000. This type of heterogeneity, reflecting countries in transition towards democracy, would not be captured using a standard fixed-effects approach.

Figure 5 in Appendix D provides the list of countries with their corresponding groups, for $G = 2$ to $G = 6$. Figures 6 and 7 show the groups of countries obtained for $G = 3$ and $G = 5$, respectively, on a world map. Examples of countries that remain non-democratic throughout the period (group 1) are China, Iran, Jordan, Syria, Egypt or Tunisia. Examples of countries that are consistently democratic (group 2) are most European countries, the US and Canada, Israel and India. Countries in transition (group 3) are Uruguay and Argentina, Brazil, Greece, and South Korea.

Accounting for a larger number of groups in estimation yields interesting results. When $G = 4$ one can clearly see two groups in transition, a medium income group which is becoming more democratic at the beginning of the sample period (group 3), and a low-income group which is becoming democratic towards the end (group 4). Group 3 contains countries such as Spain or Portugal, while Romania and South Africa belong to group 4. When allowing for $G = 5$ groups, one sees the appearance of another group whose level of democracy remains close to constant over the period, intermediate between democratic and autocratic countries. Mexico and Indonesia belong to this fifth group.

The last row of Figure 4 shows the results for $G = 6$. There we see the appearance of another intermediate group. This sixth group contains countries such as Panama, Peru and Nigeria. However, one can notice that the group-specific trends become more erratic. This pattern is accentuated when further increasing $G$ (not reported). This phenomenon reflects the fact that, as $G$ increases, some small groups tend to appear, whose time effects are poorly estimated. Table 5 in Appendix D illustrates this remark by showing the number of countries per group.

In order to interpret the groups, it is also important to compute the standard errors of the group-specific time effects, in addition to the point estimates (INCOMPLETE).

8 Additional evidence and robustness checks

This section provides additional evidence on income and democracy.

8.1 Patterns of heterogeneity

As we outlined in Section 3, the grouped fixed-effects may easily be modified to accommodate various types of heterogeneity patterns. Here we report the results for several specifications. In particular, we aim at assessing the robustness of the main results on the income effect and the group composition to modelling choices.
Figure 4: Evolution of the estimated groups, 1970 – 2000

<table>
<thead>
<tr>
<th>Trends $\hat{a}_{gt}$</th>
<th>Av. democracy</th>
<th>Av. income</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Graph of Trends" /></td>
<td><img src="image" alt="Graph of Av. democracy" /></td>
<td><img src="image" alt="Graph of Av. income" /></td>
</tr>
</tbody>
</table>

Note: See the notes to Figure 3. The left column reports the group-specific time effects $\hat{a}_{gt}$ for $G = 2$ to $G = 6$ from top to bottom. The other two columns show the group-specific averages of democracy and income, respectively. Time (1970 – 2000) is shown on the x-axis.
Table 3: Income and democracy: time-invariant group effects

<table>
<thead>
<tr>
<th>$G$</th>
<th>Objective</th>
<th>BIC</th>
<th>Income ($\frac{\theta_2}{1-\theta_1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.722</td>
<td>.049</td>
<td>.257</td>
</tr>
<tr>
<td>2</td>
<td>20.874</td>
<td>.047</td>
<td>.187</td>
</tr>
<tr>
<td>3</td>
<td>19.225</td>
<td>.045</td>
<td>.129</td>
</tr>
<tr>
<td>4</td>
<td>18.649</td>
<td>.044</td>
<td>.093</td>
</tr>
<tr>
<td>5</td>
<td>18.350</td>
<td>.044</td>
<td>.059</td>
</tr>
<tr>
<td>6</td>
<td>18.137</td>
<td>.043</td>
<td>.030</td>
</tr>
<tr>
<td>7</td>
<td>17.940</td>
<td>.043</td>
<td>.033</td>
</tr>
<tr>
<td>8*</td>
<td>17.826</td>
<td>.043</td>
<td>.024</td>
</tr>
<tr>
<td>9</td>
<td>17.761</td>
<td>.043</td>
<td>-.002</td>
</tr>
<tr>
<td>10</td>
<td>17.721</td>
<td>.043</td>
<td>.024</td>
</tr>
<tr>
<td>20</td>
<td>17.558</td>
<td>.044</td>
<td>-.038</td>
</tr>
<tr>
<td>20</td>
<td>17.517</td>
<td>-</td>
<td>-.044</td>
</tr>
</tbody>
</table>

Note: See the notes to Table 2. The standard errors of the cumulative income effect were computed using a residual bootstrap procedure (200 replications).

Table 3 shows the estimation results for the time-invariant, grouped structure of unobserved country heterogeneity of model (13). We see that the point estimate decreases when adding more groups, until becoming insignificant from zero when $G \geq 6$. Not surprisingly, we also see that the time-invariant GFE estimate is very similar to the fixed-effects estimate when $G = 20$ groups are allowed in estimation.

We also estimated model (16), which allows for two layers of heterogeneity. Imposing that each group has the same number of subgroups (e.g., $H_1 = H_2 = \ldots = H_G$) we found more variation in income estimates than in Table 2. For the (BIC-) optimal number of groups $G = 9, H = 2$, however,

---

The iterative algorithm with 10,000 random starting values was used. Contrary to Table 2, our experiments on simulated data suggested that the large-$T$ estimate of the asymptotic variance of the estimator became somewhat inaccurate for moderate to large values of $G$. Instead, in Table 3 we report estimates using a simple bootstrap strategy that we describe in the next subsection.
Table 4: Monte Carlo exercise

<table>
<thead>
<tr>
<th>G</th>
<th>Estimates</th>
<th>Monte Carlo (1)</th>
<th>Monte Carlo (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\theta_1$</td>
<td>$\theta_2$</td>
<td>$\theta_1$</td>
</tr>
<tr>
<td>3</td>
<td>.407</td>
<td>.089</td>
<td>.151</td>
</tr>
<tr>
<td></td>
<td>(.052)</td>
<td>(.011)</td>
<td>(.014)</td>
</tr>
<tr>
<td>5</td>
<td>.255</td>
<td>.079</td>
<td>.107</td>
</tr>
<tr>
<td></td>
<td>(.051)</td>
<td>(.010)</td>
<td>(.010)</td>
</tr>
</tbody>
</table>

Note: Mean and standard deviation of GFE estimates across 1000 samples. Errors are i.i.d normal (1) or resampled in a clustered wild bootstrap fashion, see the text (2). The left part of the table shows the original GFE estimates together with their standard errors.

We found a similar estimate as in our benchmark specification, the cumulative effect of income being .112 with a standard error of .014. Figures 8 and 9 show the results for $G = 3$, $H_1 = 5$, $H_2 = 2$ and $H_3 = 2$. For that specification the cumulative income effect is .068 with a standard error .008.

8.2 A Monte Carlo exercise

To assess the performance of the grouped fixed-effects estimator on these data, we performed the following simulation exercise. Based on GFE parameter estimates (for a given value of $G$) we simulated 1000 data samples using two different strategies. In the first strategy (1) we generated errors $v_{it}^{(s)}$ from an i.i.d normal distribution with a variance that we estimated on the data. In the second strategy (2) we set $v_{it}^{(s)} = \nu_i^{(s)}(y_{it} - x_{it}'\hat{\theta} - \tilde{a}_g(\tilde{g}, \tilde{a})t)$, where $\nu_i^{(s)}$ is an i.i.d standard normal shock. This second strategy is analogous to a wild bootstrap resampling scheme, clustered at the country level.

Table 4 reports the mean and standard deviation of GFE estimates across simulations, for $G = 3$ and $G = 5$. We see that the bias of the GFE estimator is generally small. The only exception is for $G = 3$ in design (2), where the autoregressive coefficient is biased downwards by 15%. However, note that in this case the implied cumulative income effect is almost unbiased. In addition, note that the Monte Carlo standard deviations are not too different from the large-$T$ standard errors estimates. Overall, the results in Table 4 suggest that the grouped fixed-effects estimator is a reliable estimation method on this type of data.\footnote{In addition, we checked that the BIC criterion effectively selects the right number of groups with high probability, for various number of groups including the optimal number in Table 2 ($G = 10$).}

We exploited this Monte Carlo exercise in two other ways. First, we computed the factor-analytic interactive fixed-effects estimator of Bai (2009). We estimated Bai’s estimator for three factors on the data generated according to the GFE model with $G = 3$ and i.i.d normal errors. In this case,
the interactive fixed-effects is consistent as \( N \) and \( T \) tends to infinity. However, we found substantial biases in the resulting estimates, with mean and standard deviations of the three parameters of interest being \(-.306 (.053), .163 (.025), \) and \(.125 (.020)\), respectively. This suggests that the parsimony of the GFE approach may result in substantial gains in precision.

Lastly, we also estimated a fixed-effects model on the data simulated according to GFE. For \( G = 3 \), the means and standard deviations of the parameters are\.291 (.067), .022 (.031), and .031 (.043), respectively. This shows that our findings are not contradictory with the results in Acemoglu et al. (2008). A fixed-effects estimator yields an insignificant income effect on a dataset where the true income effect is positive and substantial, but where unobserved country heterogeneity is time-varying.

### 8.3 Robustness checks

We have tried a number of robustness checks.

- We have included additional covariates, following Acemoglu et al. (2008), namely log population, education and age structure variables. We found similar results.

- We have used a different measure of democracy: polity IV. The results are very similar to the ones reported.

- We have used our method on the unbalanced panel dataset (122 countries, up to 9 periods). We found qualitatively similar results. Quantitatively though, estimates and group description are slightly different, as we are adding “marginal” countries. The extension to unbalanced panels and the results are to be reported in Appendix D (INCOMPLETE).

- We still have to estimate tobit model as in Benhabib et al. (2011).

**TO BE COMPLETED**

### 9 Conclusion

The grouped fixed-effects (GFE) estimator offers a flexible yet parsimonious alternative to available fixed-effects approaches for modelling unobserved heterogeneity patterns.

Though subject to an incidental parameter problem, GFE shows attractive large-\( N, T \) properties. In particular, there is no need to perform (higher-order) bias reduction. Computing the global minimum is challenging. We found the connection to data clustering and finite mixture modelling useful to devise efficient algorithms. More work in that direction is certainly needed.

The application to income and democracy evidences the usefulness of the approach in a cross-country context, where time-varying unobserved heterogeneity is very likely. Applying the method to firm or individual micro panel datasets, where \textit{ex-ante} grouping is often difficult, may be of interest.
We plan to pursue this line of work in two main directions. Firstly, the grouped fixed-effects approach is not limited to linear models. We are particularly interested in dynamic discrete choice models, where discrete modelling of unobserved heterogeneity may be attractive (Kasahara and Shimotsu, 2009, Browning and Carro, 2011).

Another direction of research concerns the possibility that the grouped structure is only approximately satisfied in the data. Bester and Hansen (2010) study the trade-off that arises when increasing $G$ in this case, between misspecification bias and incidental parameter bias. Studying this trade-off in a context where the data grouping is unknown and needs to be estimated seems an interesting avenue for future work.

References

[1] TO BE COMPLETED


APPENDIX

A Proofs

A.1 Proof of Theorem 1

Let us define:

\[ \hat{M}(\theta, \alpha) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - x'_{it} \theta - \alpha_{it})^2. \]  

Note that the GFE estimator of \((\theta, \alpha)\) minimizes \(\hat{M}(\cdot)\) on the parameter space.

Note that:

\[ \hat{M}(\theta, \alpha) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (v_{it} + x'_{it} (\theta^0 - \theta) + \alpha_{it}^0 - \alpha_{it})^2. \]

Lastly, we also define the following auxiliary objective function:

\[ \tilde{M}(\theta, \alpha) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x'_{it} (\theta^0 - \theta) + \alpha_{it}^0 - \alpha_{it})^2 + \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it}^2. \]

Step 1. We start by showing that \(\hat{M}(\cdot) - \tilde{M}(\cdot)\) becomes uniformly small asymptotically.

We have, for all \((\theta, \alpha) \in \Theta \times \mathcal{A}^{NT}:

\[ \hat{M}(\theta, \alpha) - \tilde{M}(\theta, \alpha) = \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} (x'_{it} (\theta^0 - \theta) + \alpha_{it}^0 - \alpha_{it}) \]

\[ = -\frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} \alpha_{it} + o_p(1), \]

where we have used Assumptions 1a, 1b, and 1e. Note that the \(o_p(1)\) term of the right-hand side of the last equality is uniform over the parameter space.

Now, from the fact that \(\alpha_{it} = a_{git}\):

\[ \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} \alpha_{it} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} a_{git} \]

\[ = \sum_{g=1}^{G} \left[ \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} 1\{g_{it} = g\} v_{it} a_{git} \right] \]

\[ = \sum_{g=1}^{G} \left[ \frac{1}{T} \sum_{t=1}^{T} a_{gt} \left( \frac{1}{N} \sum_{i=1}^{N} 1\{g_{it} = g\} v_{it} \right) \right]. \]

Moreover, using the Cauchy-Schwarz inequality we have, for all \(g \in \{1, \ldots, G\}:

\[ \left( \frac{1}{T} \sum_{t=1}^{T} a_{gt} \left( \frac{1}{N} \sum_{i=1}^{N} 1\{g_{it} = g\} v_{it} \right) \right)^2 \leq \left( \frac{1}{T} \sum_{t=1}^{T} a_{gt}^2 \right) \times \left( \frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{N} \sum_{i=1}^{N} 1\{g_{it} = g\} v_{it} \right)^2 \right), \]

where, by Assumption 1a:

\[ \frac{1}{T} \sum_{t=1}^{T} a_{gt}^2 = o_p(1), \]

41
uniformly over the parameter space.

Lastly, we have:

\[
\frac{1}{T} \sum_{t=1}^{T} \left( \frac{1}{N} \sum_{i=1}^{N} 1\{g_i = g\} v_{it} \right)^2 = \frac{1}{T N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} 1\{g_i = g\} 1\{g_j = g\} \sum_{t=1}^{T} v_{it} v_{jt} \leq \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| \frac{1}{T} \sum_{t=1}^{T} v_{it} v_{jt} \right|,
\]

which is \(o_p(1)\) by Assumption 1c.

This shows that \(\tilde{M}(\cdot) - \hat{M}(\cdot)\) tends to zero as \(N\) and \(T\) tend to infinity, uniformly on the parameter space.

**Step 2.** We now bound \(\tilde{M}(\cdot)\) from below. To see this, note that:

\[
\tilde{M}(\theta, \alpha) - \hat{M}(\theta^0, \alpha^0) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it}' (\theta^0 - \theta) + \alpha^0_{it} - \alpha_{it})^2.
\]

We need the following result.

**Lemma A1** Suppose that Assumption 1d holds. Then there exists a constant \(c > 0\) such that, for all \((\theta, \alpha)\) in the parameter space:

\[
\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it}' (\theta^0 - \theta) + \alpha^0_{it} - \alpha_{it})^2 \geq c \|\theta^0 - \theta\|^2 + o_p(1).
\]

**Proof.** For all \((\theta, \alpha)\) in the parameter space there exists a partition \(\{g_i\}\) of \(\{1, ..., N\}\) into \(G\) groups such that \(\alpha_{it} = a_{g_{i.t}}\). Let us denote:

\[
\Sigma_\theta(\{g_i\}) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it}' - x_{g_{i.g.i.t}}') (x_{it}' - x_{g_{i.g.i.t}}').
\]

We have, using Assumption 1e:

\[
\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it}' (\theta^0 - \theta) + \alpha^0_{it} - \alpha_{it})^2 \geq (\theta^0 - \theta)' \Sigma_\theta(\{g_i\}) (\theta^0 - \theta)
\]

\[
\geq (\theta^0 - \theta)' \left( \inf_{\{g_i\} \in \mathcal{G}_{N,G}} \Sigma_\theta(\{g_i\}) \right) \|\theta^0 - \theta\|^2 + o_p(1),
\]

where \(c\) is given in Assumption 1e. \(\blacksquare\)

Lemma A1 thus implies that, as \(N\) and \(T\) tend to infinity:

\[
\tilde{M}(\theta, \alpha) - \hat{M}(\theta^0, \alpha^0) \geq c \|\theta^0 - \theta\|^2 + o_p(1).
\]

(A2)

**Consistency of \(\hat{\theta}\).** By uniform convergence (step 1) we have:

\[
\tilde{M}(\hat{\theta}, \hat{\alpha}) = \tilde{M}(\hat{\theta}, \hat{\alpha}) - \hat{M}(\hat{\theta}, \hat{\alpha}) + \tilde{M}(\hat{\theta}, \hat{\alpha}) = \tilde{M}(\hat{\theta}, \hat{\alpha}) + o_p(1).
\]
By the definition of the GFE estimator, it thus follows that:
\[
\hat{M} \left( \hat{\theta}, \hat{\alpha} \right) \leq \hat{M} \left( \theta^0, \alpha^0 \right) + o_p(1).
\]

So, applying again the uniform convergence convergence result of step 1:
\[
\hat{M} \left( \hat{\theta}, \hat{\alpha} \right) \leq \hat{M} \left( \theta^0, \alpha^0 \right) + o_p(1).
\]

Then, (A2) implies that \( \hat{\theta} \) is consistent for \( \theta^0 \).

**Consistency of \( \hat{\alpha} \).** Lastly, note that, by the Cauchy-Schwartz inequality:
\[
\left| \hat{M} \left( \hat{\theta}, \hat{\alpha} \right) - \hat{M} \left( \theta^0, \alpha^0 \right) \right| = \left| \frac{1}{N+T} \sum_{i=1}^{N+T} x'_{i \theta} \left( \theta^0 - \hat{\theta} \right) \right| + \left| \frac{1}{N+T} \sum_{i=1}^{N+T} x'_{i \theta} \left( \theta^0 - \hat{\theta} \right) \right|
\]
\[
\leq \left( \theta^0 - \hat{\theta} \right) \left( \frac{1}{N+T} \sum_{i=1}^{N+T} x_{i \theta} x'_{i \theta} \left( \theta^0 - \hat{\theta} \right) \right) + \left( 4 \sup_{a \in \mathcal{A}} \left| a_{i \theta} \right| \right) \left( \frac{1}{N+T} \sum_{i=1}^{N+T} \left| x_{i \theta} \right| \right),
\]
which is \( o_p(1) \) by Assumptions 1a and 1e and by consistency of \( \hat{\theta} \).

Combining with (A3) we obtain:
\[
\hat{M} \left( \theta^0, \alpha^0 \right) \leq \hat{M} \left( \theta^0, \alpha^0 \right) + o_p(1),
\]
from which it follows that:
\[
\frac{1}{N+T} \sum_{i=1}^{N+T} \left( \alpha_{it}^0 - \hat{\alpha}_{it} \right)^2 = o_p(1).
\]

This completes the proof of Theorem 1.

**A.2 Proof of Corollary 1**

We study the two terms in the \( \max \{ , \} \) in turn.

**Step 1.** Let \( g \in \{ 1, \ldots, G \} \). We have:
\[
\frac{1}{N+T} \sum_{i=1}^{N} \left( \min_{g \in \{ 1, \ldots, G \}} \sum_{t=1}^{T} \mathbb{1} \left( g_{i}^0 = g \right) \left( \hat{a}_{i g t} - a_{i g t}^0 \right)^2 \right) = \left( \frac{1}{N} \sum_{i=1}^{N} \mathbb{1} \left( g_{i}^0 = g \right) \right) \left( \min_{g \in \{ 1, \ldots, G \}} \frac{1}{T} \sum_{t=1}^{T} \left( \hat{a}_{i g t} - a_{i g t}^0 \right)^2 \right).
\]

By Assumption 2a it is thus enough to show that, as \( N \) and \( T \) tend to infinity:
\[
\frac{1}{N+T} \sum_{i=1}^{N} \left( \min_{g \in \{ 1, \ldots, G \}} \sum_{t=1}^{T} \mathbb{1} \left( g_{i}^0 = g \right) \left( \hat{a}_{i g t} - a_{i g t}^0 \right)^2 \right) \rightarrow 0.
\]

Now:
\[
\frac{1}{N+T} \sum_{i=1}^{N} \left( \min_{g \in \{ 1, \ldots, G \}} \sum_{t=1}^{T} \mathbb{1} \left( g_{i}^0 = g \right) \left( \hat{a}_{i g t} - a_{i g t}^0 \right)^2 \right) \leq \frac{1}{N+T} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbb{1} \left( g_{i}^0 = g \right) \left( \hat{a}_{i g t} - a_{i g t}^0 \right)^2 \quad \rightarrow \quad \frac{1}{N+T} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \hat{a}_{i t} - a_{i t}^0 \right)^2,
\]
43
which tends to zero in probability by Theorem 1.

**Step 2.** Step 1 has shown that for all \( \varepsilon > 0 \) and for all \( g \in \{1, ..., G\} \) there exists some \( \sigma(g) \in \{1, ..., G\} \) such that, with probability approaching one:

\[
\frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{\sigma(g)t} - a^0_{gt})^2 < \varepsilon.
\]

Moreover, by the triangular inequality we have, for all \( g \neq \tilde{g} \):

\[
\left( \frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{g_{\sigma(g)t}} - \hat{a}_{g_{\tilde{g}t}})^2 \right)^{\frac{1}{2}} \geq \left( \frac{1}{T} \sum_{t=1}^{T} (a^0_{gt} - a^0_{\tilde{g}t})^2 \right)^{\frac{1}{2}} - 2\varepsilon,
\]

where \( \frac{1}{T} \sum_{t=1}^{T} (a^0_{gt} - a^0_{\tilde{g}t})^2 \) is bounded from below as \( T \) tends to infinity by Assumption 2b. So, by choosing \( \varepsilon \) small enough we see that \( \sigma(g) \) is unique. This implies that \( \sigma : \{1, ..., G\} \rightarrow \{1, ..., G\} \) is one-to-one and admits a well-defined inverse \( \sigma^{-1} \).

Hence, for all \( \tilde{g} \in \{1, ..., G\} \) and as \( T \) tends to infinity:

\[
\min_{g \in \{1, ..., G\}} \frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{g_{\sigma(g)t}} - a^0_{gt})^2 \leq \frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{g_{\sigma^{-1}(\tilde{g})t}} - a^0_{\sigma^{-1}(\tilde{g})t})^2 \xrightarrow{p} 0.
\]

This completes the proof.

### A.3 Proof of Theorem 2

To prove the result we will show that:

\[
\hat{\theta} - \tilde{\theta} = O_p \left(T^{-\frac{1}{2}}\right), \tag{A4}
\]

where \( q \) is given in assumption (3).

This will imply that:

\[
\sqrt{NT} \left(\hat{\theta} - \tilde{\theta}\right) = O_p \left(\sqrt{\frac{N}{T^{\frac{1}{2}}}}\right).
\]

The theorem will then follow by the Mann-Wald lemma under the assumption that \( \frac{N}{T^{\frac{1}{4}}} \rightarrow 0. \)

**Step 1.** For any \( \eta > 0 \) we define as \( \mathcal{N}_\eta \) the set of parameters \( (\theta, a) \in \Theta \times A^{GT} \) that satisfy \( \|\theta - \theta^0\|^2 < \eta \) and \( \frac{1}{T} \sum_{t=1}^{T} (a_{gt} - a^0_{gt})^2 < \eta \) for all \( g \in \{1, ..., G\} \).

The first step of the proof consists in showing that, for \( \eta > 0 \) small enough and as \( N \) and \( T \) tend to infinity:

\[
\sup_{(\theta, a) \in \mathcal{N}_\eta} \left| \hat{Q}(\theta, a) - \tilde{Q}(\theta, a) \right| = O_p \left(T^{-\frac{1}{2}}\right), \tag{A5}
\]

where \( q \) is given in Assumption 3.

To proceed, note that for all \( (\theta, a) \in \Theta \times A^{GT} \):

\[
\hat{Q}(\theta, a) - \tilde{Q}(\theta, a) = \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (a^*_{gt} - a^*_{\tilde{g}t}) \left( y_{it} - x_{it}' \theta - \frac{a^*_{(\theta, a)t} + a_{gt}}{2} \right).
\]

44
By the Cauchy-Schwartz inequality:

\[(\hat{Q}(\theta, a) - \bar{Q}(\theta, a))^2 \leq \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (a_{gt}^i - a_{\hat{g}_i(\theta, a)t})^2 \leq \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - x_{it}^\theta - \frac{a_{\hat{g}_i(\theta, a)t}^i + a_{g_t}^i}{2}\right)^2,\]

where the second term on the right-hand side is uniformly bounded by Assumptions 1a and 1e.

Now:

\[\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (a_{gt}^i - a_{\hat{g}_i(\theta, a)t})^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} 1\{\hat{g}_i(\theta, a) \neq g_0^i\} (a_{gt}^i - a_{\hat{g}_i(\theta, a)t})^2 \leq \left(4 \sup_{\alpha \in \mathcal{A}} a^2\right) \times \frac{1}{N} \sum_{i=1}^{N} 1\{\hat{g}_i(\theta, a) \neq g_0^i\}.

So, by Assumption 1a it is enough to bound the average misclassification error. Now, from the definition of \(\hat{g}_i(\cdot)\) we have, for all \(g \in \{1,...,G\}^2\):

\[1\{\hat{g}_i(\theta, a) = g\} \leq 1\left\{\sum_{t=1}^{T} (y_{it} - x_{it}^\theta - a_{gt})^2 \leq \sum_{t=1}^{T} (y_{it} - x_{it}^\theta - a_{g_t}^i)^2\right\}.

So:

\[\frac{1}{N} \sum_{i=1}^{N} 1\{\hat{g}_i(\theta, a) \neq g_0^i\} = \sum_{g=1}^{G} \frac{1}{N} \sum_{i=1}^{N} 1\{g_0^i \neq g\} 1\{\hat{g}_i(\theta, a) = g\} \leq \sum_{g=1}^{G} \frac{1}{N} \sum_{i=1}^{N} 1\{g_0^i \neq g\} \left\{\sum_{t=1}^{T} (y_{it} - x_{it}^\theta - a_{gt})^2 \leq \sum_{t=1}^{T} (y_{it} - x_{it}^\theta - a_{g_t}^i)^2\right\}_{Z_{g}(\theta, a)}.

Next, note that by the Markov inequality we have for all \(M > 0\):

\[\Pr\left(\sup_{(\theta, a) \in \mathcal{N}_q} \frac{1}{N} \sum_{i=1}^{N} Z_{ig}(\theta, a) > MT^{-\frac{1}{2}}\right) \leq \mathbb{E}\left[\frac{\sup_{(\theta, a) \in \mathcal{N}_q} \frac{1}{N} \sum_{i=1}^{N} Z_{ig}(\theta, a)}{MT^{-\frac{1}{2}}}\right] \leq \mathbb{E}\left[\frac{1}{N} \sum_{i=1}^{N} \sup_{(\theta, a) \in \mathcal{N}_q} \Pr(Z_{ig}(\theta, a) = 1) \right].

The result is thus a consequence of the following lemma.

**Lemma A2** Let Assumptions 1a and 3a-3c hold. Then for small enough \(\eta > 0\):

\[\frac{1}{N} \sum_{i=1}^{N} \sup_{(\theta, a) \in \mathcal{N}_q} \Pr(Z_{ig}(\theta, a) = 1) = O_p(1).

**Proof.** We have:

\[\Pr(Z_{ig}(\theta, a) = 1) \leq \max_{\tilde{g} \neq g} \Pr\left(\sum_{t=1}^{T} (y_{it} - x_{it}^\theta - a_{gt})^2 \leq \sum_{t=1}^{T} (y_{it} - x_{it}^\theta - a_{g_t}^i)^2 \mid g_0^i = \tilde{g}\right)
\]

\[= \max_{\tilde{g} \neq g} \Pr\left(\sum_{t=1}^{T} (a_{g_t} - a_{gt}) (y_{it} - x_{it}^\theta - \frac{a_{gt} + a_{g_t}}{2}) \leq 0 \mid g_0^i = \tilde{g}\right)
\]

\[= \max_{\tilde{g} \neq g} \Pr\left(\sum_{t=1}^{T} (a_{g_t} - a_{gt}) (v_{it} + x_{it}^\theta (\theta^0 - \theta) + e_{g_t}^0 - a_{gt} + a_{g_t}^0) \leq 0 \mid g_0^i = \tilde{g}\right).
\]
Let $\tilde{g} \in \{1, \ldots, G\}$ such that $\tilde{g} \neq g$. We start by showing that, for $\eta > 0$ small enough:
\[
\sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) \left( a_{\tilde{g}t}^0 - \frac{a_{gt} + a_{\tilde{g}t}}{2} \right) \geq \frac{T}{2} c_{g, \tilde{g}}. \tag{A6}
\]
To see this, let:
\[
A_T \equiv \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) \left( a_{\tilde{g}t}^0 - \frac{a_{gt} + a_{\tilde{g}t}}{2} \right) - \frac{1}{2} \sum_{t=1}^{T} (a_{\tilde{g}t}^0 - a_{gt}^0)^2.
\]
We have, by simple rearrangement:
\[
A_T = \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) \left( a_{\tilde{g}t}^0 - \frac{a_{gt} + a_{\tilde{g}t}}{2} \right) - \sum_{t=1}^{T} (a_{\tilde{g}t}^0 - a_{gt}^0) \left( a_{\tilde{g}t}^0 - \frac{a_{gt}^0 + a_{\tilde{g}t}^0}{2} \right)
= \sum_{t=1}^{T} a_{\tilde{g}t}^0 (a_{\tilde{g}t}^0 - a_{gt}^0 - a_{gt} + a_{\tilde{g}t}) + \frac{1}{2T} \sum_{t=1}^{T} ([a_{\tilde{g}t}^0]^2 - [a_{gt}^0]^2 - [a_{gt}^0]^2 + [a_{gt}]^2).
\]
It thus follows from the Cauchy-Schwartz inequality and Assumption 1a that, for $(\theta, a) \in N_\eta$:
\[
|A_T| \leq TC \sqrt{\eta},
\]
where $C$ is independent of $\eta$ and $T$. Hence (A6) follows from choosing $\eta > 0$ small enough and using Assumption 3a.

The next step in the proof is to note that:
\[
\left| \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) x_{it} (\theta^0 - \theta) \right| \leq T \left( 2 \sup_{a_t \in A} |a_t| \right) \sup \|x_{it}\| \|\theta^0 - \theta\| \leq T \tilde{C} \sqrt{\eta}, \tag{A7}
\]
where $\tilde{C}$ is independent of $\eta$ and $T$, and where we have used Assumptions 1a and 3d.

Combining (A6) and (A7) and taking $\eta \leq \left( \frac{c_{g, \tilde{g}}}{2C} \right)^2$, it is thus sufficient to bound:
\[
\Pr \left( \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) v_{it} \leq -\frac{T}{4} c_{g, \tilde{g}} \right),
\]
where we omit the conditioning on $g_t^0 = \tilde{g}$ for conciseness.

Now, using Assumption 1a (to ensure that $a_{\tilde{g}t} - a_{gt}$ is bounded in $\mathbb{R}$), Assumptions 3b and 3c and the Markov inequality, we have:
\[
\Pr \left( \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) v_{it} \leq -\frac{T}{4} c_{g, \tilde{g}} \right) \leq \Pr \left[ \left( \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) v_{it} \right)^q \geq \left( \frac{T}{4} c_{g, \tilde{g}} \right)^q \right]
\leq \mathbb{E} \left( \left( \sum_{t=1}^{T} (a_{\tilde{g}t} - a_{gt}) v_{it} \right)^q \right)
\leq \frac{B_i}{\left( \frac{T}{4} c_{g, \tilde{g}} \right)^q} T^{-\frac{q}{2}},
\]
where $q$ and $B_i$ are given in Assumption 3.

Lastly, by Assumption 3c we have: \( \frac{1}{N} \sum_{i=1}^{N} B_i = O_p(1) \). This completes the proof of Lemma A2.

\[\blacksquare\]
Step 2. Note that by Assumptions 1a-1e and Assumptions 2a-2b we have that \( \hat{\theta} \overset{p}{\to} \theta^0 \), and:

\[
\frac{1}{T} \sum_{t=1}^{T} (\hat{a}_{\sigma(g)t} - \tilde{a}_{gt})^2 \overset{p}{\to} 0,
\]

where the permutation \( \sigma \) is as in the proof of Corollary 1. By simple relabelling of the elements of \( \hat{a} \) we may take \( \sigma(g) = g \).

As a consequence, for all \( \eta > 0 \) and as \( N \) and \( T \) tend to infinity:

\[
\Pr \left( \left( \hat{\theta}, \hat{a} \right) \notin N_{\eta} \right) \to 0,
\]

where \( N_{\eta} \) was defined in step 1.

Together with (A8), this result implies that, for small enough \( \eta > 0 \):

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) = O_p \left( T^{-\frac{4}{7}} \right). \tag{A8}
\]

Note that we also have, asymptotically in \( N \) and \( T \):

\[
\Pr \left( \left( \tilde{\theta}, \tilde{a} \right) \notin N_{\eta} \right) \to 0,
\]

so that:

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) = O_p \left( T^{-\frac{4}{7}} \right).
\]

Next, note that, by the definition of \( (\tilde{\theta}, \tilde{a}) \):

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) \geq 0.
\]

Moreover, using the above and the definition of \( (\tilde{\theta}, \tilde{a}) \):

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) = \tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) + O_p(T^{-\frac{4}{7}}) \leq O_p \left( T^{-\frac{4}{7}} \right).
\]

It thus follows that:

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) = O_p \left( T^{-\frac{4}{7}} \right).
\]

Now, we have:

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) = 2 \frac{N}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( x_{it} \left( \hat{\theta} - \tilde{\theta} \right) + \tilde{a}_{g_{it}} - \tilde{a}_{g_{it}} \right) \left( y_{it} - x_{it} \left( \hat{\theta} + \tilde{\theta} \right) \right) \left( \frac{1}{2} \right) - \frac{\tilde{a}_{g_{it}} + \tilde{a}_{g_{it}}}{2} \left( \frac{1}{2} \right)
\]

\[
= \left( \hat{\theta} - \tilde{\theta} \right)^T 2 \frac{N}{T} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \left( y_{it} - x_{it} \left( \frac{1}{2} \right) - \frac{\tilde{a}_{g_{it}} + \tilde{a}_{g_{it}}}{2} \right)
\]

\[
+ \frac{1}{T} \sum_{g=1}^{G} \sum_{t=1}^{T} \left( \tilde{a}_{gt} - \tilde{a}_{gt} \right) \frac{2}{N} \sum_{i=1}^{N} 1_{g_{it} = g} \left( y_{it} - x_{it} \left( \frac{1}{2} \right) - \frac{\tilde{a}_{g_{it}} + \tilde{a}_{g_{it}}}{2} \right). \tag{A9}
\]
Note that, as \((\hat{\theta}, \hat{a})\) is a least squares estimator, the following empirical moment restrictions are satisfied:

\[
\begin{align*}
\frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} (y_{it} - x_{it}' \hat{\theta} - \tilde{a}_{gt}) &= 0 \\
\frac{2}{N} \sum_{i=1}^{N} 1\{g_{i}^{0} = g\} (y_{it} - x_{it}' \hat{\theta} - \tilde{a}_{gt}) &= 0, \text{ for all } (g, t).
\end{align*}
\]

Combining with (A9) yields:

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) = \left( \hat{\theta} - \tilde{\theta} \right)' \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} \left( x_{it}' \left( \frac{\hat{\theta} - \tilde{\theta}}{2} \right) + \sum_{g=1}^{G} 1\{g_{i}^{0} = g\} \left( \frac{\tilde{a}_{gt} - \hat{a}_{gt}}{2} \right) \right) \\
+ \frac{1}{T} \sum_{g=1}^{G} (\tilde{a}_{gt} - \hat{a}_{gt}) \sum_{i=1}^{N} 1\{g_{i}^{0} = g\} \left( \left( x_{it}' \frac{\hat{\theta} - \tilde{\theta}}{2} \right) + \frac{\tilde{a}_{gt} - \hat{a}_{gt}}{2} \right)
\]

(A10)

\[
\geq \left( \hat{\theta} - \tilde{\theta} \right)' \left( \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - \bar{x}_{gt, t}) (x_{it} - \bar{x}_{gt, t})' \right) \left( \hat{\theta} - \tilde{\theta} \right).
\]

It thus follows that:

\[
\tilde{Q} (\hat{\theta}, \hat{a}) - \tilde{Q} (\hat{\theta}, \hat{a}) \geq \hat{\rho} \| \hat{\theta} - \tilde{\theta} \|^2,
\]

where \(\hat{\rho}\) tends in probability to the minimum eigenvalue \(\rho > 0\) of \(\Sigma_{\theta}\), which is positive by Assumption 3d.

This shows (A4) and ends the proof of Theorem 2.

### A.4 Proof of Corollary 2

Note that:

\[
\left\| \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \sum_{g=1}^{G} 1\{g_{i}^{0} = g\} \left( \frac{\tilde{a}_{gt} - \hat{a}_{gt}}{2} \right) \right) \right\| \leq \frac{1}{T} \sum_{g=1}^{G} \| \tilde{a}_{g} - \hat{a}_{g} \| \left( \sum_{t=1}^{T} 1\{g_{i}^{0} = g\} \| x_{it} \| \right)^{2} \leq \frac{\lambda_1}{\sqrt{T}} \sum_{g=1}^{G} \| \tilde{a}_{g} - \hat{a}_{g} \|
\]

where we have used the Cauchy-Schwartz inequality and Assumptions 1e and 3d.

Using (A10) it follows that:

\[
\frac{1}{T} \sum_{g=1}^{G} \| \tilde{a}_{g} - \hat{a}_{g} \|^2 \leq \left| \tilde{Q} \left( \hat{\theta}, \hat{a} \right) - \tilde{Q} \left( \hat{\theta}, \hat{a} \right) \right| + C_{2} \| \hat{\theta} - \tilde{\theta} \|^2 + \frac{C_{1}}{\sqrt{T}} \| \hat{\theta} - \tilde{\theta} \| \sum_{g=1}^{G} \| \tilde{a}_{g} - \hat{a}_{g} \|.\n\]

Hence, for all \(g \in \{1, ..., G\}:

\[
\frac{1}{T} \| \tilde{a}_{g} - \hat{a}_{g} \|^2 \leq \left| \tilde{Q} \left( \hat{\theta}, \hat{a} \right) - \tilde{Q} \left( \hat{\theta}, \hat{a} \right) \right| + C_{2} \| \hat{\theta} - \tilde{\theta} \|^2 + \frac{C_{1}}{\sqrt{T}} \| \hat{\theta} - \tilde{\theta} \| \| \tilde{a}_{g} - \hat{a}_{g} \|.\n\]

This is a second-order polynomial inequality in \(\frac{1}{\sqrt{T}} \| \tilde{a}_{g} - \hat{a}_{g} \|\), the solution of which satisfies:

\[
\frac{1}{\sqrt{T}} \| \tilde{a}_{g} - \hat{a}_{g} \| \leq O_{p} \left( T^{-\frac{1}{2}} \right),
\]

48
where we have used that, by the proof of Theorem 2, \( |\tilde{Q}(\tilde{\theta}, \tilde{\alpha}) - \tilde{Q}(\tilde{\theta}, \tilde{\alpha})| = O_p(T^{-\frac{1}{2}}) \), and \( \|\tilde{\theta} - \hat{\theta}\| = O_p(T^{-\frac{1}{2}}) \).

This implies that:

\[
\frac{1}{\sqrt{T}} \|\tilde{a}_g - a^0_g\| \leq \frac{1}{\sqrt{T}} \|\tilde{a}_g - a_g\| + \frac{1}{\sqrt{T}} \|\tilde{a}_g - a^0_g\| = O_p(T^{-\frac{1}{2}}) + O_p\left(\frac{1}{\sqrt{N}}\right),
\]

where the first term is \( O_p(T^{-\frac{1}{2}}) \) by the above and the second term is \( O_p\left(\frac{1}{\sqrt{N}}\right) \) by (33).

So, as \( N \) and \( T \) tend to infinity such that \( \frac{N}{T^4} \) tends to zero we have:

\[
\frac{1}{\sqrt{T}} \|\tilde{a}_g - a^0_g\| = O_p\left(\frac{1}{\sqrt{N}}\right).
\]

The result then comes from taking squares.

A.5 Proof of Corollary 3

The proof closely follows that of Theorem 2. The key difference is the following lemma, which replaces Lemma A2. To state the lemma it is useful to define the following quantity:

\[
Z_{ig}^{(\pi)} = 1\{g_i = g\} \{\sum_{t=1}^T \left( y_{it} - x_{it}'\theta - a_{gt} \right)^2 - C \ln \pi_{ig} \leq \sum_{t=1}^T \left( y_{it} - x_{it}'\theta - a^0_{gt} \right) \leq C \ln \pi_{ig} \}
\]

**Lemma A3** Let the assumptions of Lemma A2 hold, and let \( \pi_i \) satisfy Assumption 4. Then for small enough \( \eta > 0 \):

\[
\frac{1}{N} \sum_{i=1}^N \sup_{(\theta, a) \in \mathcal{N}_\eta} \Pr \left( Z_{ig}^{(\pi)}(\theta, a) = 1 \right) = O_p(1).
\]

**Proof.** The only difference with the proof of Lemma A2 is that we now bound, see (A7):

\[
\left| \sum_{t=1}^T (a_{gt} - a_g) x_{it}' (\theta^0 - \theta) - C \ln \pi_{ig} + C \ln \pi_{ig} \right| \leq T \left( 2 \sup_{a_t \in A} |a_t| \right) \sup \|x_{it}\| \|\theta^0 - \theta\| + 2C \ln \varepsilon
\]

\[
\leq T^2 \sqrt{\eta} + 2C \ln \varepsilon
\]

for \( T \) large enough, where we have used Assumption 4.

The rest of the proof is identical to the one of Lemma A2. \( \Box \)
A.6 Proof of Proposition 1

Let $G \geq G^0$. We start by noting that:

$$\Delta^{(G)} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x'_{it} \hat{\theta}^{(G)} - \hat{\alpha}_{it}^{(G)} \right)^2 - \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( y_{it} - x'_{it} \theta^{(0)} - \alpha_{it}^{(0)} \right)^2$$

$$= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( v_{it} + x'_{it} \left( \theta^{(0)} - \hat{\theta}^{(G)} \right) + \alpha_{it}^{(0)} - \hat{\alpha}_{it}^{(G)} \right)^2 - \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it}^2$$

$$= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( x'_{it} \left( \theta^{(0)} - \hat{\theta}^{(G)} \right) + \alpha_{it}^{(0)} - \hat{\alpha}_{it}^{(G)} \right)^2$$

$$+ \frac{2}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} \left( x'_{it} \left( \theta^{(0)} - \hat{\theta}^{(G)} \right) + \alpha_{it}^{(0)} - \hat{\alpha}_{it}^{(G)} \right).$$

We will show that $\Delta^{(G)} = O_p(1/N)$. To proceed, note that $\|\hat{\theta}^{(G)} - \theta^{(0)}\| = O_p(1/\sqrt{NT})$ by (40). Moreover:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \hat{\alpha}_{it}^{(G)} - \alpha_{it}^{(0)} \right)^2 = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{y=1}^{G^0} \sum_{\tilde{y}=1}^{G} 1 \{ g_i^{(0)} = g \} 1 \{ \tilde{g}_i = \tilde{y} \} \left( \hat{\alpha}_{it}^{(G)} - a_{gt}^{(0)} \right)^2$$

$$\leq \frac{1}{T} \sum_{t=1}^{T} \sum_{y=1}^{G^0} \sum_{\tilde{y}=1}^{G} \left( \hat{\alpha}_{it}^{(G)} - a_{gt}^{(0)} \right)^2,$$

which is $O_p(1/N)$ by (41).

It easily follows from these rates of convergence and Assumption 1e that:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( x'_{it} \left( \theta^{(0)} - \hat{\theta}^{(G)} \right) + \alpha_{it}^{(0)} - \hat{\alpha}_{it}^{(G)} \right)^2 = O_p \left( \frac{1}{N} \right).$$

Moreover, from Assumptions 1b, 3b and 3d we have:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} x_{it} = O_p \left( \frac{1}{\sqrt{NT}} \right).$$

This implies that:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} \left( x'_{it} \left( \theta^{(0)} - \hat{\theta}^{(G)} \right) \right) = O_p \left( \frac{1}{NT} \right).$$

As for the last term in $\Delta^{(G)}$ we have:

$$\frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} \left( \hat{\alpha}_{it}^{(G)} - \alpha_{it}^{(0)} \right) = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{y=1}^{G^0} \sum_{\tilde{y}=1}^{G} 1 \{ g_i^{(0)} = g \} 1 \{ \tilde{g}_i = \tilde{y} \} \left( \hat{\alpha}_{it}^{(G)} - a_{gt}^{(0)} \right),$$

where $\hat{g}_i$ is a shorthand for $\hat{g}_i \left( \hat{\theta}^{(G)}, \hat{\alpha}^{(G)} \right)$.

Moreover, for all $g, \tilde{g}$ we have, by the Cauchy-Schwartz inequality:

$$\left( \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} v_{it} 1 \{ g_i^{(0)} = g \} 1 \{ \hat{g}_i = \tilde{g} \} \left( \hat{\alpha}_{it}^{(G)} - a_{gt}^{(0)} \right) \right)^2 \leq \frac{1}{N^2 T^2} \left( \sum_{t=1}^{T} \left( \hat{\alpha}_{it}^{(G)} - a_{gt}^{(0)} \right)^2 \right) \times \ldots \times \left[ \sum_{t=1}^{T} \left( \sum_{i=1}^{N} 1 \{ g_i^{(0)} = g \} 1 \{ \hat{g}_i = \tilde{g} \} v_{it} \right)^2 \right].$$

50
Now, by (41):

\[
\frac{1}{T} \sum_{t=1}^{T} \left( \hat{a}_{yt}^{(G)} - a_{yt}^{0} \right)^2 = O_p \left( \frac{1}{N} \right).
\]

In addition:

\[
\frac{1}{N^2 T} \sum_{t=1}^{T} \left( \sum_{i=1}^{N} 1 \{ \hat{g}_i = g \} \{ \hat{g}_i = \hat{g} \} v_{it} \right)^2 \leq \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{1}{T} \sum_{t=1}^{T} v_{it} v_{jt} \right),
\]

which is \(O_P(1/N)\) by Assumption 1c. This shows that \(\Delta^{(G)} = O_P(1/N)\).

Hence, for all \(G > G^0\):

\[
\hat{I}(G) - \hat{I}(G^0) = \Delta^{(G)} - \Delta^{(G^0)} + (G - G^0) h_{NT}
\]

\[
= O_p(1/N) + (G - G^0) h_{NT}.
\]

It thus follows that:

\[
N \left( \hat{I}(G) - \hat{I}(G^0) \right) = O_p(1) + (G - G^0) Nh_{NT},
\]

where the \(O_p(1)\) term is negligible as \(Nh_{NT} \xrightarrow{N,T \to \infty} \infty\).

This shows that \(\hat{G} \leq G^0\) with probability one.

**B Additional statistical results**

**B.1 Fixed-\(T\) bias**

In this subsection of the appendix we study some properties of the GFE estimator in an asymptotic where \(T\) is kept fixed as \(N\) tends to infinity.

Let us consider the probability limit of the GFE estimator in model (8) as \(N\) tends to infinity, for fixed \(T\):

\[
\left( \hat{\theta}, \hat{a} \right) \xrightarrow{p} \operatorname{argmin}_{(\theta, a) \in \Theta \times A} \varlimsup_{N \to \infty} \hat{Q}(\theta, a), \tag{B11}
\]

where \(\hat{Q}(\cdot)\) is given by (A1). In order to formally establish (B11), consistency conditions for models with nonsmooth objective functions are needed (e.g., Chapter 7 in Newey and McFadden, 1994). Pollard (1981) provides a rigorous proof in the special case with no covariates, in which case model (8) is the kmeans model.\(^{32}\)

Note that, because of group labelling, the statement concerning \(\hat{a}\) should be understood relative to the Hausdorff distance.\(^{33}\)

However, the fixed-\(T\) probability limit in (B11) does not coincide with \((\theta^0, a^0)\) in general. The reason is that there always remains a misclassification bias as \(N\) tends to infinity when \(T\) is fixed, because \(\hat{\theta}_i\) differs from \(g_{0i}^\theta\) even in the limit. To understand the nature of the fixed-\(T\) bias we shall work with the moment restrictions that arise from the minimization problem in (B11). At the true value \((\theta^0, a^0)\), the moments are:

\[
- \frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( y_{it} - x_{it} \theta^0 - a_{0i}^{\theta} \right) \right] = b_{T\theta}^0, \tag{B12}
\]

\(^{32}\)An important condition to establish convergence in probability in (B11) is that the minimum be unique. Note that this fixed-\(T\) identification condition was not needed to prove Theorems 1 and 2.

\(^{33}\)Pollard (1982) provides conditions under which \(\hat{a}\) is root-\(N\) consistent and asymptotically normal for its probability limit, in the model without covariates. Here we focus only on bias, and do not discuss fixed-\(T\) inference more generally.
and, for all \( g \in \{1, \ldots, G\} \) and all \( t \in \{1, \ldots, T\} \):

\[
-\frac{2}{T} \mathbb{E} \left[ \mathbf{1}\{\tilde{g}_i^0 = g\} \left( y_{it} - x_{it}'\theta^0 - a_{g_{it}}^0 \right) \right] \equiv b_{T,a_i}^0. \tag{B13}
\]

The rest of this section consists in characterizing the bias of the moment restrictions in the special case where \( v_{it} \) are i.i.d normal. We start by providing a general expression.

**Proposition B1** The first-order conditions of the minimization problem on the right-hand side of (B11) are, evaluated at \((\theta^0, a^0)\):

\[
-\frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( y_{it} - x_{it}'\theta^0 - a_{g_{it}}^0 \right) \right] = -\frac{2}{T} \sum_{t=1}^{T} \sum_{g=1}^{G} \sum_{g \neq g} (a_{g_{it}}^0 - a_{g_{it}}^0) \mathbb{E} \left[ x_{it} \mathbf{1}\{g_i^0 = \tilde{g}\} \mathbf{1}\{\tilde{g}_i^0 = g\} \right], \tag{B14}
\]

and, for all \( g \in \{1, \ldots, G\} \) and all \( t \in \{1, \ldots, T\} \):

\[
-\frac{2}{T} \mathbb{E} \left[ \mathbf{1}\{\tilde{g}_i^0 = g\} \left( y_{it} - x_{it}'\theta^0 - a_{g_{it}}^0 \right) \right] = -\frac{2}{T} \sum_{g \neq g} (a_{g_{it}}^0 - a_{g_{it}}^0) \mathbb{E} \left[ \mathbf{1}\{g_i^0 = \tilde{g}\} \mathbf{1}\{\tilde{g}_i^0 = g\} \right] \\
+ \frac{2}{T} \sum_{g \neq g} \mathbb{E} \left[ \mathbf{1}\{g_i^0 = g\} \mathbf{1}\{\tilde{g}_i^0 = \tilde{g}\} v_{it} \right], \tag{B15}
\]

where:

\[
\mathbf{1}\{\tilde{g}_i^0 = g\} = \mathbf{1} \left\{ \sum_{i=1}^{T} (a_{hi_{it}}^0 - a_{g_{it}}^0) \left( v_{it} + a_{g_{it}}^0 - a_{g_{it}}^0 \right) \leq 0 \quad \text{for all } h \in \{1, \ldots, G\}, \quad h \neq g \right\}. \tag{B16}
\]

and where \( \tilde{g}_i^0 \equiv \tilde{g}_i(\theta^0, a^0) \) is given by (18).

**Proof.**

We start by noting that, for all \( g \in \{1, \ldots, G\} \) and for all \((\theta, a) \in \Theta \times \mathcal{A}^{GT}\), the set of vectors \((y_i', x_i')\) that satisfies \( \tilde{g}_i(\theta, a) = g \) forms a convex polyhedron in \( \mathbb{R}^{T(K+1)} \). Assuming that the boundary of that region has zero measure,\(^{34}\) it follows from the arguments of Lemma A in Pollard (1982) that the first-order condition of the right-hand side of (B11) are:

\[
-\frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( y_{it} - x_{it}'\tilde{\theta} - a_{g_{it}}^0 \right) \right] = 0, \\
-\frac{2}{T} \mathbb{E} \left[ \mathbf{1}\{\tilde{g}_i(\theta, a) = g\} \left( y_{it} - x_{it}'\tilde{\theta} - a_{g_{it}}^0 \right) \right] = 0 \quad \text{for all } (g, t),
\]

where \((\tilde{\theta}, a)\) are the fixed-\(T\) pseudo-true parameter values given by (B11).

Next, we have:

\[
-\frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( y_{it} - x_{it}'\theta^0 - a_{g_{it}}^0 \right) \right] = -\frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( v_{it} + a_{g_{it}}^0 - a_{g_{it}}^0 \right) \right] \\
= -\frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( a_{g_{it}}^0 - a_{g_{it}}^0 \right) \right],
\]

\(^{34}\)A sufficient condition for this is that the conditional distribution of \( y_{i} \) given \( x_{i} \) be absolutely continuous.
where we have used Assumption 1b.

So:

\[-\frac{2}{T} \mathbb{E} \left[ \sum_{t=1}^{T} x_{it} \left( y_{it} - x_{it}' \theta_0 - a_{g_{it}}^0 \right) \right] = -\frac{2}{T} \sum_{t=1}^{T} \sum_{g=1}^{G} \sum_{g' \neq g} \mathbb{E} \left[ x_{it} 1 \{ g_i^0 = \tilde{g} \} 1 \{ g_i^0 = g \} \left( a_{g_{it}}^0 - a_{g'_{it}}^0 \right) \right] = -\frac{2}{T} \sum_{t=1}^{T} \sum_{g=1}^{G} \sum_{g' \neq g} \left( a_{g_{it}}^0 - a_{g'_{it}}^0 \right) \mathbb{E} \left[ x_{it} 1 \{ g_i^0 = \tilde{g} \} 1 \{ g_i^0 = g \} \right].\]

Next we have: for all \( g \in \{1,...,G\} \) and all \( t \in \{1,...,T\} \):

\[-\frac{2}{T} \mathbb{E} \left[ 1 \{ \tilde{g}_i^0 = g \} \left( y_{it} - x_{it}' \theta_0 - a_{g_{it}}^0 \right) \right] = \sum_{g' \neq g} \mathbb{E} \left[ 1 \{ g_i^0 = \tilde{g} \} 1 \{ g_i^0 = g \} \left( v_{it} + a_{g_{it}}^0 - a_{g'_{it}}^0 \right) \right] - \frac{2}{T} \mathbb{E} \left[ 1 \{ g_i^0 = \tilde{g} \} \right].\]

Now by Assumption 1b we have:

\[\mathbb{E} \left[ 1 \{ g_i^0 = g \} v_{it} \right] = 0.\]

So:

\[\mathbb{E} \left[ 1 \{ g_i^0 = g \} 1 \{ \tilde{g}_i^0 = g \} v_{it} \right] = -\sum_{g' \neq g} \mathbb{E} \left[ 1 \{ g_i^0 = g \} 1 \{ \tilde{g}_i^0 = \tilde{g} \} v_{it} \right].\]

It thus follows that:

\[-\frac{2}{T} \mathbb{E} \left[ 1 \{ \tilde{g}_i^0 = g \} \left( y_{it} - x_{it}' \theta_0 - a_{g_{it}}^0 \right) \right] = -\frac{2}{T} \sum_{g' \neq g} \left( a_{g_{it}}^0 - a_{g'_{it}}^0 \right) \mathbb{E} \left[ 1 \{ g_i^0 = \tilde{g} \} 1 \{ g_i^0 = g \} \right] - \frac{2}{T} \sum_{g' \neq g} \mathbb{E} \left[ 1 \{ g_i^0 = \tilde{g} \} \right].\]

Lastly, we have:

\[1 \{ \tilde{g}_i^0 = g \} = \begin{cases} \sum_{t=1}^{T} \left( y_{it} - x_{it}' \theta_0 - a_{g_{it}}^0 \right)^2 \leq \sum_{t=1}^{T} \left( y_{it} - x_{it}' \theta_0 - a_{g_{it}}^0 \right)^2 & \text{for all } h \in \{1,...,G\}, \ h \neq g \\ \sum_{t=1}^{T} \left( v_{it} + a_{g_{it}}^0 - a_{g_{it}}^0 \right)^2 \leq \sum_{t=1}^{T} \left( v_{it} + a_{g_{it}}^0 - a_{g_{it}}^0 \right)^2 & \text{for all } h \in \{1,...,G\}, \ h \neq g \\ \sum_{t=1}^{T} \left( a_{g_{it}}^0 - a_{g_{it}}^0 \right) \left( v_{it} + a_{g_{it}}^0 - a_{g_{it}}^0 \right) \leq 0 & \text{for all } h \in \{1,...,G\}, \ h \neq g \end{cases}.\]

Proposition B1 shows that the bias of the moment restrictions depends on the probability of events of the type \( g_i^0 = g, \ g_i^0 = \tilde{g}, \ g_i^0 = g, \ g_i^0 = \tilde{g} \). This makes it clear that the GFE bias depends on group misclassification.

The next result provides an explicit expression for the bias under normality. Some notation is needed. Let \( C_{tg}^0 \) be the \((G - 1) \times T\) matrix with \((h,t)\)-element \( (a_{ht}^0 - a_{g_{ht}}^0) \). Let also \( a_{tg}^0 \) be the \((G - 1) \times 1\) vector with \(h\)-element \( \sum_{t=1}^{T} (a_{ht}^0 - a_{g_{ht}}^0) \left( a_{g_{ht}}^0 - a_{g_{ht}}^0 \right) \). Finally, let \( C_{tg}^0 \) denotes the \( t\)-column of the matrix \( C_{tg}^0 \).

53
We will denote as $\Phi_{G-1}(:,\Sigma)$ the cumulative distribution function of the $(G-1)$-dimensional normal with zero mean and variance-covariance matrix $\Sigma$. In case where $\Sigma$ is singular we define $\Phi_{G-1}(:,\Sigma) \equiv \lim_{\varepsilon \to 0} \Phi_{G-1}(:,\Sigma + \varepsilon I_{G-1})$. We will also denote as $E_{G-1}(:,\Sigma)$ the mean of the truncated $(G-1)$-dimensional normal distribution with zero mean and variance $\Sigma$, using the same convention in case $\Sigma$ is singular.

**Corollary 4** Suppose that $v_{it}$ are i.i.d $\mathcal{N}(0,\sigma^2)$. Then we have:

$$b^0_{Tg} = -\frac{2}{T} \sum_{t=1}^{T} \sum_{g'=1}^{G} (a^0_{g't} - a^0_{g't}) \mathbb{E} \left[ x_{it} 1 \{ g^0_i = \tilde{g} \} \right] \Phi_{G-1} \left( -d^0_{Tg}; \sigma^2 C^0_{Tg} (C^0_{Tg})' \right), \quad (B17)$$

and, for all $g \in \{1,...,G\}$ and all $t \in \{1,...,T\}$:

$$b^0_{Tg}= -\frac{2}{T} \sum_{g'=g}^{G} (a^0_{g't} - a^0_{g't}) \mathbb{E} \left[ 1 \{ g^0_i = \tilde{g} \} \right] \Phi_{G-1} \left( -d^0_{Tg}; \sigma^2 C^0_{Tg} (C^0_{Tg})' \right)$$

$$-\frac{2}{T} \sum_{g' \neq g} \mathbb{E} \left[ 1 \{ g^0_i = \tilde{g} \} \right] (c^0_{g't})' \left[ C^0_{Tg} (C^0_{Tg})' \right]^{-1} E_{G-1} \left( -d^0_{Tg}; \sigma^2 C^0_{Tg} (C^0_{Tg})' \right)$$

$$+ \frac{2}{T} \sum_{g' \neq g} \mathbb{E} \left[ 1 \{ g^0_i = g \} \right] (c^0_{g't})' \left[ C^0_{Tg} (C^0_{Tg})' \right]^{-1} E_{G-1} \left( -d^0_{Tg}; \sigma^2 C^0_{Tg} (C^0_{Tg})' \right). \quad (B18)$$

Corollary 4 is a simple consequence of the following lemma.

**Lemma B4** Under the conditions of Proposition 4 we have, for all $g \neq \tilde{g}$ in $\{1,2\}$:

$$\mathbb{E} \left[ 1 \{ \tilde{g}^0_i = g \} \right] g^0_i = \tilde{g}, x_i = \Phi_{G-1} \left( -d^0_{Tg}; \sigma^2 C^0_{Tg} (C^0_{Tg})' \right), \quad (B19)$$

$$\mathbb{E} \left[ v_{it} 1 \{ \tilde{g}^0_i = g \} \right] g^0_i = \tilde{g}, x_i = (c^0_{Tg})' \left[ C^0_{Tg} (C^0_{Tg})' \right]^{-1} E_{G-1} \left( -d^0_{Tg}; \sigma^2 C^0_{Tg} (C^0_{Tg})' \right). \quad (B20)$$

**Proof.**

Suppose that $g^0_i = \tilde{g} \neq g$. By (B16) we have:

$$1 \{ \tilde{g}^0_i = g \} = 1 \left\{ \sum_{t=1}^{T} (a^0_{ht} - a^0_{g't}) v_{it} \leq - \sum_{t=1}^{T} (a^0_{ht} - a^0_{g't}) \left( a^0_{g't} - \frac{a^0_{ht} + a^0_{g't}}{2} \right) \text{ for all } h \neq g \right\}. \quad (B19)$$

(B19) follows immediately.

Next, we have:

$$\mathbb{E} \left[ v_{it} 1 \{ \tilde{g}^0_i = g \} \right] g^0_i = \tilde{g}, x_i = \mathbb{E} \left[ v_{it} 1 \{ C^0_{Tg} v_i \leq -d^0_{Tg} \} \right] | g^0_i = \tilde{g}, x_i = (c^0_{Tg})' \left[ C^0_{Tg} (C^0_{Tg})' \right]^{-1} \mathbb{E} \left[ C^0_{Tg} v_i 1 \{ C^0_{Tg} v_i \leq -d^0_{Tg} \} \right] | g^0_i = \tilde{g}, x_i,$$

and (B20) follows.

As an example, when $G = 2$ the expressions in Corollary 4 take the following simple form:

$$b^0_{Tg} = -\frac{2}{T} \sum_{t=1}^{T} \sum_{g=1}^{G} \sum_{g' \neq g} (a^0_{g't} - a^0_{g't}) \mathbb{E} \left[ x_{it} 1 \{ g^0_i = \tilde{g} \} \right] \Phi \left( -\frac{1}{2\sigma} \left( \sum_{t=1}^{T} (a^0_{g't} - a^0_{g't})^2 \right)^{\frac{1}{2}} \right) \quad (B21)$$
and, for all \((g,t)\):

\[
\begin{align*}
\theta_{T_{a,t}}^0 &= -\frac{2}{T} \sum_{\tilde{g} \neq g} (a_{gt}^0 - a_{\tilde{g}t}^0) \mathbb{E} \left[ 1\{g_0^t = \tilde{g}\} \right] \Phi \left( -\frac{1}{2\sigma} \left( \sum_{t=1}^{T} (a_{gt}^0 - a_{\tilde{g}t}^0)^2 \right)^{\frac{1}{2}} \right) \\
&\quad + \frac{2}{T} \sum_{\tilde{g} \neq g} \mathbb{E} \left[ 1\{g_0^t = \tilde{g}\} \right] \frac{\sigma (a_{gt}^0 - a_{\tilde{g}t}^0)}{\left( \sum_{t=1}^{T} (a_{gt}^0 - a_{\tilde{g}t}^0)^2 \right)^{\frac{1}{2}}} \phi \left( -\frac{1}{2\sigma} \left( \sum_{t=1}^{T} (a_{gt}^0 - a_{\tilde{g}t}^0)^2 \right)^{\frac{1}{2}} \right) \\
&\quad - \frac{2}{T} \sum_{\tilde{g} \neq g} \mathbb{E} \left[ 1\{g_0^t = g\} \right] \frac{\sigma (a_{gt}^0 - a_{\tilde{g}t}^0)}{\left( \sum_{t=1}^{T} (a_{gt}^0 - a_{\tilde{g}t}^0)^2 \right)^{\frac{1}{2}}} \phi \left( -\frac{1}{2\sigma} \left( \sum_{t=1}^{T} (a_{gt}^0 - a_{\tilde{g}t}^0)^2 \right)^{\frac{1}{2}} \right),
\end{align*}
\]

(B22)

where \(\Phi(\cdot)\) and \(\phi(\cdot)\) denote the normal cdf and pdf, respectively.

It is easy to see that the expressions in (B21) and (B22) tend exponentially fast to zero as \(T\) tends to infinity, provided Assumption 2b is satisfied.

### B.2 Consistency of \(\hat{a}\) when \(G > G^0\)

In this subsection we take \(G > G^0\), and we aim at establishing consistency of \(\hat{a}\) as \(N\) and \(T\) tend to infinity. We shall make the following assumption.

**Assumption B1 (consistency of \(\hat{a}\) when \(G > G^0\))**

For all \(\tilde{g} \in \{1, \ldots, G\}\):

\[
\operatorname{plim}_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} 1 \left\{ \hat{g}_i \left( \hat{\theta}, \hat{a} \right) = \tilde{g} \right\} > 0.
\]

Assumption B1 is a high-level assumption that requires that none of the \(G\) estimated groups is empty in the limit. We then have the following result.

**Corollary 5 (consistency of \(\hat{a}\) when \(G > G^0\))**

Let \(G > G^0\), and let the assumptions of Corollary 1 hold. Let also Assumption B1 hold. Then, as \(N\) and \(T\) tend to infinity:

\[
\max\left\{ \max_{g \in \{1, \ldots, G^0\}} \left( \min_{\tilde{g} \in \{1, \ldots, G\}} \frac{1}{T} \sum_{t=1}^{T} (\tilde{a}_{gt}^0 - a_{\tilde{g}t}^0)^2 \right), \max_{\tilde{g} \in \{1, \ldots, G\}} \left( \min_{g \in \{1, \ldots, G^0\}} \frac{1}{T} \sum_{t=1}^{T} (\tilde{a}_{gt}^0 - a_{gt}^0)^2 \right) \right\} \to 0.
\]

**Proof.** Step 1 of the proof of Corollary 1 remains valid. Now, note that step 1 only relied on Assumption 2a. When interchanging \((G^0, g^0_i)\) and \((G, \hat{g}_i \left( \hat{\theta}, \hat{a} \right))\) the analogous assumption is Assumption B1, which is assumed to hold true. Hence the second term in the \(\max\{\cdot, \cdot\}\) may be bounded in exactly the same way as the first term. This ends the proof.

### B.3 Misspecification of the number of groups: the case \(G^0 = 1, \ G = 2\)

**TO BE COMPLETED**

### C Markov Chain Monte Carlo algorithms

**INCOMPLETE**
**Gibbs sampling.** The Gibbs sampler for mixture models (Diebolt and Robert, 1994) relies on data augmentation, introducing the latent groups $g_1, \ldots, g_N$ as additional parameters to sample from.

**Algorithm 2 (Gibbs sampling)**

1. Start with initial parameter values $(\theta^{(0)}, a^{(0)}) \in \Theta \times A^G$.
   
   Set $j = 0$.

2. Compute for all $i \in \{1, \ldots, N\}$:
   
   $$g_i^{(j+1)} = \text{argmin}_{g \in \{1, \ldots, G\}} \sum_{t=1}^T \left( y_{it} - x'_{it} \theta^{(j)} - a_i^{(j)} g_i \right)^2.$$  

3. Draw $(\theta^{(j+1)}, a^{(j+1)})$ from:
   
   $$p^{(j+1)}(\theta, a | y, x, g_1^{(j+1)}, \ldots, g_N^{(j+1)}) \propto h(\theta, a) \exp \left( - \sum_{i=1}^N \sum_{t=1}^T \left( y_{it} - x'_{it} \theta - a_i^{(j+1)} g_i \right)^2 \right). \quad (C23)$$

4. Stop the Markov Chain when a stationary distribution has been reached.

In this algorithm step 2 is deterministic. This is because the group probabilities are left unrestricted. Note also that step 3 is conditional on the group values $g_i^{(j)}$, so it simply amounts to drawing from the posterior distribution of the normal linear model. At each iteration $j$, it may happen that one of the groups is empty, i.e. that $g_i^{(j+1)} \neq g$ for all $j$. In that case $a_g$ is drawn from its prior distribution (e.g., a uniform distribution of $A$).

**Metropolis-Hastings.** TO BE COMPLETED

**D Additional estimation results**
Figure 5: Groups of countries

<table>
<thead>
<tr>
<th>Country</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
<th>Group 5</th>
<th>Group 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burundi</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Benin</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central African</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>China</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Côte d’Ivoire</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Congo, Rep.</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algeria</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Egypt, Arab Rep.</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gabon</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ghana</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Guinea</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greece</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Honduras</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iran</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jordan</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kenya</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Madagascar</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mexico</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mali</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malawi</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Malawi</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Niger</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nigeria</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nepal</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pakistan</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portugal</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Republic of Chad</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romania</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian Federation</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senegal</td>
<td></td>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Singapore</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>South Africa</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turkey</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uruguay</td>
<td></td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: See the notes to Figure 3. The first to last columns show the group classification for $G = 2$ to $G = 6$ groups.
Figure 6: World map.

- **Low Democracy**
- **Transition**
- **High Democracy**
- **No data**
Table 5: Number of countries per group

<table>
<thead>
<tr>
<th></th>
<th>$g = 1$</th>
<th>$g = 2$</th>
<th>$g = 3$</th>
<th>$g = 4$</th>
<th>$g = 5$</th>
<th>$g = 6$</th>
<th>$g = 7$</th>
<th>$g = 8$</th>
<th>$g = 9$</th>
<th>$g = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G = 1$</td>
<td>90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 2$</td>
<td>41</td>
<td>49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 3$</td>
<td>28</td>
<td>24</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 4$</td>
<td>26</td>
<td>13</td>
<td>18</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 5$</td>
<td>21</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 6$</td>
<td>22</td>
<td>10</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 7$</td>
<td>21</td>
<td>3</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>11</td>
<td>29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 8$</td>
<td>20</td>
<td>7</td>
<td>6</td>
<td>2</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G = 9$</td>
<td>20</td>
<td>9</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>$G = 10$</td>
<td>20</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td>4</td>
<td>29</td>
</tr>
</tbody>
</table>

*Note: See notes to Figure 3.*

Figure 8: Two layers of heterogeneity

*Note: Two-layer model with $G = 3$, $H_1 = 5$, $H_2 = 2$, and $H_3 = 2$.***
Figure 9: The world, $G = 3, H_1 = 5, H_2 = 2, H_3 = 2$.