SUPPLEMENT TO "GROUPED PATTERNS OF HETEROGENEITY IN PANEL DATA"

(*Econometrica*, Vol. 83, No. 3, May 2015, 1147–1184)

BY STÉPHANE BONHOMME AND ELENA MANRESA

THIS SUPPLEMENT IS DIVIDED into seven sections. In Section S.1, we describe the computational algorithms. In Section S.2, we deal with inference, both from a large-N, T perspective and from a large-N, fixed-T perspective. In Section S.3, we treat the issues of misspecification of the number of groups G and its choice. In Section S.4, we study two extensions of the baseline model, which allow for unit-specific heterogeneity and for group-specific coefficients, respectively. In Section S.5, we deal with several other issues, including the connection with mixture models, and how to incorporate prior information in estimation. In Section S.6, we report the results of a simulation study. Lastly, in Section S.7, we show a number of additional results related to the empirical application.

S.1. COMPUTATION

In this section, we provide details on the two computational algorithms, and we illustrate their performance in a numerical exercise.

S.1.1. *Algorithms*

The Simple Iterative Algorithm

Algorithm 1 described in the paper is a clustering algorithm. Indeed, it coincides with the well-known *kmeans* algorithm (Forgy (1965)) in the special case where there are no covariates in the model (i.e., when $\theta = 0$). In this case, (4) boils down to the standard minimum sum-of-squares partitioning problem:

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

In geometric terms, (S.1) amounts to finding a collection of "centers" $\alpha_1, \alpha_2, \ldots, \alpha_G$ in $\mathbb{R}^T$ such that the sum of the Euclidean distances between y_i and the closest center α_g is minimum. Due to its relevance in many different fields (such as astronomy, genetics, or psychology), this problem has been extensively studied in operations research and computer science (Steinley (2006)).

A drawback of Algorithm 1 is its dependence on the chosen starting values. One way to overcome this problem is to choose many random starting values, and then select the solution that yields the lowest objective. In the numerical experiments reported below and the empirical application, we use the following method to generate starting values:

 DOI: 10.3982/ECTA11319

1. Draw $\theta^{(0)}$ from some prespecified distribution supported on Θ.
2. Draw G units $i_1, i_2, \ldots, i_G$ in $\{1, \ldots, N\}$ at random, and set

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

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

A More Efficient Algorithm

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

ALGORITHM 2—Variable Neighborhood Search:

1. *Let* $(\theta, \alpha) \in \Theta \times \mathcal{A}^{GT}$ *be some starting value*.
Perform one assignment step of Algorithm 1 *and obtain an initial grouping* γ_{init}.
Set $\text{iter}_{\max}$ *and* $\text{neigh}_{\max}$ *to some desired values*.
Set $j = 0$.
Set $\gamma^* = \gamma_{\text{init}}$.
2. *Set* $n = 1$.
3. (*Neighborhood jump*) *Relocate* n *randomly selected units to* n *randomly selected groups, and obtain a new grouping* γ'.
Perform one update step of Algorithm 1 *and obtain new parameter values* (θ', α').
4. *Set* $(\theta^{(0)}, \alpha^{(0)}) = (\theta', \alpha')$, *and apply Algorithm* 1.
5. (*Local search*) *Starting from the grouping* $\gamma = \{g_1, \ldots, g_N\}$ *obtained in Step* 4, *systematically check all reassignments of units* $i \in \{1, \ldots, N\}$ *to groups* $g \in \{1, \ldots, G\}$ (*for* $g \neq g_i$), *updating* g_i *when the objective function decreases*; *stop when no further reassignment improves the objective function*.
Let the resulting grouping be γ''.
6. *If the objective function using* γ'' *improves relative to the one using* γ^*, *then set* $\gamma^* = \gamma''$ *and go to Step* 2; *otherwise, set* $n = n + 1$ *and go to Step* 7.

7. *If* $n \leq neigh_{\max}$, *then go to Step 3*; *otherwise go to Step 8*.
8. *Set* $j = j + 1$. *If* $j > iter_{\max}$, *then Stop*; *otherwise go to Step 2*.

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

Choice of Tuning Parameters

Algorithm 2 depends on two parameters set by the researcher: the maximum neighborhood size $neigh_{\max}$, and a maximum number of iterations $iter_{\max}$. The algorithm may also be run using different starting parameter values, even though the choice of starting values tends to matter much less than in the case of Algorithm 1. Denoting as N_s the number of starting values, Algorithm 2 is thus indexed by $(N_s; neigh_{\max}; iter_{\max})$. The parameter $iter_{\max}$ measures the length of the computation for a given starting value, and may be interpreted as a stopping rule. The parameter $neigh_{\max}$ represents the number of neighborhoods evaluated during the search. We follow previous implementation (see Brusco and Steinley (2007)) and set $neigh_{\max} = 10$. We also set $iter_{\max} = 10$ and $N_s = 10$ in our main estimation exercises.[1] A practical rule of thumb for choosing the tuning parameters is to check that different starting values tend to yield the exact same solution.

S.1.2. *Numerical Performance*

Tables S.I and S.II show the value of the final objective corresponding to different computational methods, on the cross-country panel data set that we use in the empirical application. The data set has dimensions $N = 90$, $T = 7$, and two covariates (including a lagged outcome). We show the value of the objective and computation time for both algorithms when $G = 2, 3$, and 10. We show the results for the first 30 countries, the first 60 countries (alphabetically ordered), and all 90 countries in the data set.

[1]In several of the exercises that we performed, these choices resulted in prohibitive computation times. As a result, the bootstrapped standard errors in Figure 1 in the paper, as well as the Monte Carlo estimates in Tables S.III and S.VI, were computed using $(N_s; neigh_{\max}; iter_{\max}) = (5, 10, 5)$. The estimates of bootstrapped standard errors in the Monte Carlo exercise in Tables S.IV and S.VII were computed using Algorithm 1 with 1,000 starting values.

TABLE S.I

NUMERICAL PERFORMANCE ($G = 2, 3$)[a]

	Algorithm 1 (1,000)		Algorithm 2 (10; 10; 10)		
	Value	Time	Value	Time	Exact Value
			$G = 2$		
$N = 30$	6.159	0.6	6.159	2.1	6.159*
$N = 60$	13.209	0.9	13.209	7.6	13.209*
$N = 90$	19.846	1.3	19.846	18.2	19.846*
			$G = 3$		
$N = 30$	4.913	0.6	4.913	6.1	4.913*
$N = 60$	10.934	1.1	10.934	16.7	10.934**
$N = 90$	16.598	1.7	16.598	38.4	16.598**

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

Table S.I suggests that the simple iterative algorithm performs well when the number of groups is small. Algorithms 1 and 2 yield the same solution (i.e., the same objective and optimal grouping) in all configurations of the data. In contrast, Table S.II shows that Algorithm 2 improves on Algorithm 1 when the number of groups increases. When $G = 10$ and $N = 30$, running the iterative algorithm using 1,000 starting values yields a higher value for the objective function than when using Algorithm 2. When all $N = 90$ countries are included in Table S.II, even 1,000,000 different starting values and a running time of approximately one hour yields a higher objective than when using Algorithm 2

TABLE S.II

NUMERICAL PERFORMANCE ($G = 10$)[a]

	Algorithm 1 (1,000)		Algorithm 1 (1,000,000)		Algorithm 2 (10; 10; 10)		Algorithm 2 (1,000; 20; 20)		
	Value	Time	Value	Time	Value	Time	Value	Time	Exact Value
$N = 30$	1.106	1.1	1.025	988.3	1.025	48.3	1.025	10,872.2	1.025**
$N = 60$	4.373	2.0	4.255	1,729.5	4.255	116.4	4.255	28,301.9	N/A
$N = 90$	8.035	3.4	7.762	3,235.6	7.749	228.4	7.749	132,555.7	7.749***

[a]See note to Table S.I. In the "exact" column, *** refers to Aloise, Hansen, and Liberti's (2012) exact column generation algorithm for given $\widehat{\theta}$.

during only four minutes of search (7.749 versus 7.762, respectively).[2] Interestingly, running Algorithm 2 during 36 hours yields exactly the same objective and grouping.

Despite these results, one concern is that even the best heuristic methods can lead to nonoptimal solutions. To assess whether the solutions of Algorithm 2 are optimal in Tables S.I and S.II, we make use of—and extend—*exact* algorithms for the minimum sum-of-squares partitioning problem. New methods have recently been proposed to compute globally optimal solutions in this challenging problem,[3] including Brusco's (2006) repetitive branch and bound algorithm, and Aloise, Hansen, and Liberti's (2012) column generation algorithm. In the "exact" columns of Tables S.I and S.II (indicated with two or three stars), we report the objective function obtained when applying one of these exact algorithms to the vector of residuals $y_{it} - x_{it}'\widehat{\theta}$, where $\widehat{\theta}$ has been computed using our best heuristic (Algorithm 2). We see that the objective and grouping coincide with the ones identified by Algorithm 2 in all cases, including when $G = 10$. This provides very encouraging evidence on the performance of our algorithm, and confirms previous evidence obtained for minimum sum-of-squares partitioning (Brusco and Steinley (2007)).

In addition, we were able to extend Brusco's (2006) repetitive branch and bound algorithm to allow for covariates.[4] Although our current implementation is limited to a small number of groups ($G = 2$ for $N \leq 90$, and $G = 3$ for $N = 30$), it yields the same solution as the one obtained using the heuristics; see the results indicated with one star in Table S.I. This formally demonstrates that our heuristic algorithm has correctly identified the global minimum in these cases.

Overall, this section suggests that the computation problem for GFE is challenging, yet not impossible, thanks to recent advances in data clustering. Our main algorithm (Algorithm 2) delivers fast and reliable estimates, and we have provided evidence that the solutions obtained are globally optimal in the data set of our empirical application. In larger data sets, the simple iterative algorithm (Algorithm 1) is a practical option.[5] Assessing the numerical perfor-

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

[3]It has been proved that problem (S.1) may be solved exactly in $O(N^{GT+1})$ operations (Inaba, Katoh, and Imai (1994)).

[4]The extension of the algorithm that allows for covariates is available as Supplemental Material.

[5]In large data sets, an alternative is to proceed in three steps: first estimate the GFE estimator on a random subsample of size $n \ll N$, yielding $(\widehat{\theta}^{(0)}, \widehat{\alpha}^{(0)})$; then classify all N units in the entire sample based on $(\widehat{\theta}^{(0)}, \widehat{\alpha}^{(0)})$; finally estimate $(\widehat{\theta}, \widehat{\alpha})$ using an OLS regression on the estimated groups, using the entire sample. Though not numerically equal to the argument of the global minimum of the GFE objective function, this three-step estimator will be asymptotically equivalent to the latter in a large-N, T perspective, under the conditions spelled out in Section 3 of the paper, provided $n \to \infty$. We thank Denis Chetverikov for pointing this out to us.

mance of the two algorithms as the dimensions of the problem increase is a natural next step.

Finally, it is worth pointing out that research on computational algorithms is still in progress. Mixed Integer Nonlinear Programming (MINLP) is an active area of research. Recent work has shown that sophisticated interior point methods can deliver exact solutions to problem (S.1) in competitive time in several large instances. While Brusco's (2006) repetitive branch and bound algorithm computed the global minimum in (S.1) in Fisher's Iris data ($N = 150$, $T = 4$) for as much as $G = 10$ groups, Du Merle, Hansen, Jaumard, and Mladenovic (2001) and more recently Aloise, Hansen, and Liberti (2012) computed exact solutions in data sets of dimensions up to $N = 2310$ and $T = 19$, for $G = 250$ groups.[6] We view exact and heuristic methods as complementary tools to compute GFE estimators.

S.2. INFERENCE

In this section, we first present estimators of the large N, T variance of the GFE estimator in model (1). Then we study the large-N, fixed-T asymptotic properties of GFE, and propose variance estimators.

S.2.1. *Large-N, T Inference*

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

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

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

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

$$\widehat{\mathrm{Var}}(\widehat{\theta}) = \frac{\widehat{\Sigma}_\theta^{-1}\widehat{\Omega}_\theta\widehat{\Sigma}_\theta^{-1}}{NT}, \tag{S.3}$$

[6]Note that the algorithm of Aloise, Hansen, and Liberti (2012) that we used in Table S.II delivered the global optimum in 1.7 seconds only.

where, denoting as $\overline{x}_{gt}$ the mean of x_{it} in group $\widehat{g}_i = g$,[7] we take

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

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

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

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

The properties of Arellano's (1987) formula in fixed-effects models as N and T tend to infinity are studied in Hansen (2007).

Finally, note that the assumptions of Corollary 1 allow for weak dependence in the cross-sectional dimension, too. However, the variance formulas (S.2)–(S.3) are generally invalid in that case. The literature provides a number of variance estimators that account for spatial and time-series dependence, and can be applied to GFE estimators. For example, when a meaningful notion of distance d_{ij} between units is available, one can construct the following estimator of Ω_θ that is robust to serial correlation and spatial correlation that diminishes with distance:

$$\widetilde{\Omega}_\theta = \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\kappa\left(\frac{d_{ij}}{d_N}\right)\widehat{v}_{it}\widehat{v}_{js}(x_{it} - \overline{x}_{\widehat{g}_i,t})(x_{js} - \overline{x}_{\widehat{g}_j,s})',$$

where κ is a kernel function, and d_N is suitably chosen as an increasing function of the sample size, as in Kelejian and Prucha (2007) and Moscone and Tosetti (2012).

Below we show numerical evidence on the finite sample performance of the estimator (S.3) of the variance of the GFE estimator. In the exercises on simulated and real data, we use variance formulas clustered at the unit (i.e., country) level. Thus, we implicitly assume away spatial dependence, while taking into account general forms of time dependence.

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

S.2.2. *Large-N, Fixed-T Inference*

S.2.2.1. *Asymptotic Distribution*

Let $(\widehat{\theta}, \widehat{\alpha})$ be the GFE estimator of (θ, α) in model (1). Let also $y_i = (y_{i1}, \ldots, y_{iT})'$ (with dimensions $T \times 1$), and $x_i = (x_{i1}, \ldots, x_{iT})'$ ($T \times K$, where $K = \dim x_{it}$). We assume that (y_i, x_i) are i.i.d. across individuals and have finite second moments. Note that, in contrast with the large-N, T analysis in the paper, here we assume random sampling across units. In addition, we assume that the solution to the following population minimization problem:

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

is unique up to relabeling. Lastly, we assume that the solution to every minimization problem of the form (S.4) but based on $\widetilde{G} < G$ groups is also unique. Then, extending the analysis of Pollard (1981) to allow for covariates, it can be shown that, as N tends to infinity with T fixed,

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

Note that, in contrast with the asymptotic analysis of Section 3 in the paper, uniqueness of the solution in (S.4) does not require the data generating process to have a grouped structure.[8]

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

$$\mathbb{E}\left[x_i'(y_i - x_i\overline{\theta} - \overline{\alpha}_{\widehat{g}_i(\overline{\theta},\overline{\alpha})})\right] = 0, \tag{S.5}$$

and

$$\mathbb{E}\left[\mathbf{1}\{\widehat{g}_i(\overline{\theta}, \overline{\alpha}) = g\}(y_i - x_i\overline{\theta} - \overline{\alpha}_g)\right] = 0 \quad \text{for all } g = 1, \ldots, G, \tag{S.6}$$

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

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

$$\sqrt{N}\begin{pmatrix}\widehat{\theta} - \overline{\theta} \\ \widehat{\alpha} - \overline{\alpha}\end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \Gamma^{-1}V\Gamma^{-1}\right), \tag{S.7}$$

[8]On the other hand, this assumption rules out purely homogeneous DGPs as soon as $T \geq 2$. To see this, suppose that y_{it} are i.i.d. standard normal, and that there are no covariates in the model. In the case $T = 2$, it can be shown that the solutions to (S.4) lie on a circle whose radius is identified, but that the precise location of the points on the circle is not.

where the $(GT + K) \times (GT + K)$ matrices V and Γ are defined below. As in Pollard's (1982) main theorem, for (S.7) to hold we assume that y_i has a continuous density given x_i, and that Γ is positive definite, in addition to the assumptions needed for consistency.

The GFE estimator $(\widehat{\theta}, \widehat{\alpha})$ is a just-identified GMM estimator based on non-smooth moment functions. V is given by

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

where

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

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

Moreover, Γ is given by

$$\Gamma = \begin{pmatrix} \Gamma_{\theta\theta} & \Gamma_{\theta 1} & \cdots & \Gamma_{\theta G} \\ \Gamma_{1\theta} & \Gamma_{11} & \cdots & \Gamma_{1G} \\ \cdots & \cdots & \cdots & \cdots \\ \Gamma_{G\theta} & \Gamma_{G1} & \cdots & \Gamma_{GG} \end{pmatrix},$$

where

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

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

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

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

The next result provides a convenient alternative expression for Γ.

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

$$S_{gh} = \big\{y \in \mathbb{R}^T, \|y - x\theta - \alpha_g\|^2 = \|y - x\theta - \alpha_h\|^2, \text{ and } \|y - x\theta - \alpha_g\|^2 \leq \|y - x\theta - \alpha_{\widetilde{h}}\|^2 \text{ for all } \widetilde{h} \neq (g, h)\big\}. \tag{S.8}$$

We denote S_{gh} as $\overline{S}_{gh}$ when evaluated at $(\overline{\theta}, \overline{\alpha})$.[9]

[9]Note that S_{gh} and $\overline{S}_{gh}$ depend on x, although we leave the dependence implicit for conciseness. Moreover, the integrals are relative to the $(T-1)$-dimensional Lebesgue measure.

We have

$$\Gamma_{\theta\theta} = \mathbb{E}\big[x_i' x_i\big] - \frac{1}{2}\sum_{g=1}^{G}\sum_{h\neq g}\mathbb{E}\bigg[\bigg(\int_{\overline{S}_{gh}} f(y|x_i)\,dy\bigg)x_i'\bigg(\frac{(\overline{\alpha}_h-\overline{\alpha}_g)(\overline{\alpha}_h-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}\bigg)x_i\bigg], \tag{S.9}$$

$$\Gamma_{\theta g} = \mathbb{E}\big[x_i'\mathbf{1}\big\{\widehat{g}_i(\overline{\theta},\overline{\alpha})=g\big\}\big] + \sum_{h\neq g}\mathbb{E}\bigg[x_i'(\overline{\alpha}_g-\overline{\alpha}_h)\bigg(\int_{\overline{S}_{gh}}\frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}f(y|x_i)\,dy\bigg)\bigg], \tag{S.10}$$

$$\Gamma_{gg} = \mathbb{E}\big[\mathbf{1}\big\{\widehat{g}_i(\overline{\theta},\overline{\alpha})=g\big\}\big]I_T - \mathbb{E}\bigg[\sum_{h\neq g}\bigg(\int_{\overline{S}_{gh}}\frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)(y-x_i\overline{\theta}-\overline{\alpha}_g)'}{\|\overline{\alpha}_h-\overline{\alpha}_g\|}f(y|x_i)\,dy\bigg)\bigg], \tag{S.11}$$

$$\Gamma_{g\widetilde{g}} = \mathbb{E}\bigg[\bigg(\int_{\overline{S}_{g\widetilde{g}}}\frac{(y-x_i\overline{\theta}-\overline{\alpha}_g)(y-x_i\overline{\theta}-\overline{\alpha}_{\widetilde{g}})'}{\|\overline{\alpha}_{\widetilde{g}}-\overline{\alpha}_g\|}f(y|x_i)\,dy\bigg)\bigg] \quad \textit{for all } \widetilde{g}\neq g. \tag{S.12}$$

For the proof see Appendix S.A.3.

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

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

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

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

$$\int_{-\infty}^{(\overline{\alpha}_1+\overline{\alpha}_2)/2}(y-\overline{\alpha}_1)f(y)\,dy=0$$

and

$$\int_{(\overline{\alpha}_1+\overline{\alpha}_2)/2}^{+\infty}(y-\overline{\alpha}_2)f(y)\,dy=0,$$

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

It is easily verified that

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

The second term in Γ represents the contribution to the variance due to observations that are at the margin between group 1 and group 2. Note that, if the DGP is given by equation (14) in the paper with $\alpha_1^0 \neq \alpha_2^0$, and denoting as ϕ the standard normal density,

$$\begin{aligned} &\left|\frac{\overline{\alpha}_2 - \overline{\alpha}_1}{4}\right| f\left(\frac{\overline{\alpha}_1 + \overline{\alpha}_2}{2}\right) \\ &= \left|\frac{\alpha_2^0 - \alpha_1^0}{4} + o_p\big(T^{-\delta}\big)\right| \\ &\quad \times \frac{\sqrt{T}}{\sigma}\left[\Pr\big(g_i^0 = 1\big)\phi\left(\frac{\sqrt{T}}{\sigma}\left(\frac{\alpha_2^0 - \alpha_1^0}{2} + o_p\big(T^{-\delta}\big)\right)\right)\right. \\ &\quad \left. + \Pr\big(g_i^0 = 2\big)\phi\left(\frac{\sqrt{T}}{\sigma}\left(\frac{\alpha_1^0 - \alpha_2^0}{2} + o_p\big(T^{-\delta}\big)\right)\right)\right], \end{aligned}$$

which tends to zero as T tends to infinity. When groups are well-separated, and under suitable tail and dependence conditions, the additional variance contribution due to group misclassification vanishes asymptotically. As a result, the large-N, fixed-T formula tends to the large-N, T formula as T tends to infinity.

S.2.2.2. *Variance Estimation*

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

Analytical Formulas. A consistent estimator of V is readily obtained as

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

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

$$\begin{aligned} \widehat{\Gamma}_{\theta\theta} &= \frac{1}{N}\sum_{i=1}^{N} x_i' x_i \\ &\quad - \frac{1}{2N}\sum_{g=1}^{G}\sum_{h\neq g}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) x_i' \left(\frac{(\widehat{\alpha}_h - \widehat{\alpha}_g)(\widehat{\alpha}_h - \widehat{\alpha}_g)'}{\|\widehat{\alpha}_h - \widehat{\alpha}_g\|}\right) x_i, \end{aligned} \tag{S.13}$$

$$\widehat{\Gamma}_{\theta g} = \frac{1}{N}\sum_{i=1}^{N} x_i' \mathbf{1}\{\widehat{g}_i(\widehat{\theta},\widehat{\alpha}) = g\} + \frac{1}{N}\sum_{h\neq g}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) x_i' (\widehat{\alpha}_g - \widehat{\alpha}_h) \frac{(y_i - x_i\widehat{\theta} - \widehat{\alpha}_g)'}{\|\widehat{\alpha}_h - \widehat{\alpha}_g\|}, \tag{S.14}$$

$$\widehat{\Gamma}_{gg} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\widehat{g}_i(\widehat{\theta},\widehat{\alpha}) = g\} I_T - \frac{1}{N}\sum_{h\neq g}\sum_{i=1}^{N} \widehat{\Delta}_{igh}(\epsilon_N) \frac{(y_i - x_i\widehat{\theta} - \widehat{\alpha}_g)(y_i - x_i\widehat{\theta} - \widehat{\alpha}_g)'}{\|\widehat{\alpha}_h - \widehat{\alpha}_g\|}, \tag{S.15}$$

$$\widehat{\Gamma}_{g\widetilde{g}} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\Delta}_{ig\widetilde{g}}(\epsilon_N) \frac{(y_i - x_i\widehat{\theta} - \widehat{\alpha}_g)(y_i - x_i\widehat{\theta} - \widehat{\alpha}_{\widetilde{g}})'}{\|\widehat{\alpha}_{\widetilde{g}} - \widehat{\alpha}_g\|} \quad \text{for all } \widetilde{g} \neq g, \tag{S.16}$$

where

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

and where $\kappa(\cdot)$ is a kernel function. Note that $\frac{\overline{\alpha}_h - \overline{\alpha}_g}{\|\overline{\alpha}_h - \overline{\alpha}_g\|}$ is the normal vector to the hypersurface $\overline{S}_{gh}$. The estimator $\widehat{\Gamma}$ is reminiscent of Powell's (1986) variance estimator for quantile regression. Similarly, $\widehat{\Gamma}$ will be consistent for Γ if $\epsilon_N \to 0$ and $\sqrt{N}\epsilon_N \to +\infty$. To implement this method, we take a Gaussian kernel $\kappa = \phi$. Optimal choice of ϵ_N exceeds the scope of this paper.[10]

Bootstrap. An alternative to the analytical formulas $\widehat{V}$ and $\widehat{\Gamma}$ is to use the bootstrap, resampling unit-specific blocks of observations (y_i, x_i) from the

[10]We experimented with the following nonadaptive rule, roughly mimicking Silverman's (1986) rule of thumb for density estimation:

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

and obtained good results on simulated and real data. This is the choice we used in Tables S.IV, S.VII, and S.XI.

original sample. Consistency of the bootstrap for the minimum sum-of-squares partitioning problem, relying on the asymptotic derivations of Pollard (1982) and the results on the bootstrap obtained by Giné and Zinn (1990), is shown in Arcones and Giné (1992). As it requires multiple optimization of the GFE objective for different samples, however, the bootstrap is computationally intensive compared to the inference approach based on analytical formulas.

S.3. UNKNOWN NUMBER OF GROUPS

The asymptotic results of Section 3 in the paper were derived under the assumption that the true number of groups G^0 was known. In this section, we relax this assumption and let G be the (possibly incorrect) number of groups postulated by the researcher.

S.3.1. *Incorrect Number of Groups: A Simple Case*

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

PROPOSITION S.2: *Let us consider the model*

$$y_{it} = x_{it}'\theta^0 + \alpha^0_{g_i^0} + v_{it}, \quad v_{it} \sim \text{i.i.d. } \mathcal{N}\left(0, \sigma^2\right), v_{it} \text{ independent of } x_{js}, \tag{S.17}$$

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

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

For the proof see Appendix S.A.4.

In this example, the data generating process is homogeneous ($G^0 = 1$), but the researcher estimates two groups ($G = 2$). The proof of Proposition S.2 shows that, asymptotically, the two estimated groups are solely based on random errors (depending on whether $\overline{v}_i \geq 0$). Given that the spurious groups

are independent of covariates, their presence does not bias the GFE estimator of θ^0. In fact, allowing for a larger number of groups than the true one in GFE estimation may be thought of as including $(G - G^0)$ irrelevant regressors—uncorrelated with the covariates of interest—in a linear regression. A similar intuition applies to interactive fixed-effects models: Moon and Weidner (2010a) showed that the asymptotic distribution of the interactive fixed-effects estimator with $G \geq G^0$ factors is identical to that of the estimator based on the correct number of factors. We conjecture that this result applies to the GFE estimator in model (1). However, a formal proof of this conjecture is beyond the scope of this paper.

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

S.3.2. *Estimating the Number of Groups*

To consistently estimate the number of groups G^0 in model (1), we rely on the connection with the analysis of large factor models and interactive fixed-effects panel data models and consider the following class of information criteria:

$$I(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\big(y_{it} - x_{it}'\widehat{\theta}^{(G)} - \widehat{\alpha}_{\widehat{g}_{it}}^{(G)}\big)^2 + Gh_{NT}, \tag{S.18}$$

where $^{(G)}$ refers to the GFE estimator with G groups, and h_{NT} is a penalty. The estimated number of groups is then

$$\widehat{G} = \underset{G\in\{1,\ldots,G_{\max}\}}{\operatorname{argmin}} I(G), \tag{S.19}$$

where $G_{\max}$ is an upper bound on G^0.

Following the arguments in Bai and Ng (2002) and Bai (2009), it can be shown that the estimated number of groups $\widehat{G}$ is consistent for G^0 if, as N and T tend to infinity, h_{NT} tends to zero and $\min(N, T)h_{NT}$ tends to infinity. The first condition ensures that $\widehat{G} \geq G^0$ with probability approaching 1, while the second condition guarantees that $\widehat{G} \leq G^0$. The availability of a known upper bound $G_{\max}$ is key in order to derive the asymptotic properties. The problem of selecting $G_{\max}$ is not considered here.

As an example, let us consider the following Bayesian Information Criterion (BIC):[11]

$$\text{BIC}(G) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\big(y_{it} - x_{it}'\widehat{\theta}^{(G)} - \widehat{\alpha}^{(G)}_{\widehat{g}_i t}\big)^2 + \widehat{\sigma}^2\frac{GT+N+K}{NT}\ln(NT), \tag{S.20}$$

where $\widehat{\sigma}^2$ is a consistent estimate of the variance of v_{it}.[12] One easily sees that the BIC estimate $\widehat{G}$ is consistent for G^0 if N and T tend to infinity at the same rate. In contrast, if T tends to infinity more slowly than N so that T/N tends to zero, the BIC criterion (S.20) implies that $\text{plim}_{N,T\to\infty}\widehat{G} \geq G^0$, but $\widehat{G}$ may be inconsistent for G^0.

S.4. EXTENSIONS OF THE BASELINE MODEL

In this section, we analyze the large-N, T properties of the GFE estimator in two models: model (5) that combines time-invariant unit-specific heterogeneity with time-varying grouped patterns, and model (7) that allows for group-specific coefficients.

S.4.1. *Extension 1: Unit-Specific Heterogeneity*

Consider model (5), and define

$$\big(\widehat{\theta}^{FE}, \widehat{\mu}^{FE}, \widehat{\gamma}^{FE}\big) = \underset{(\theta,\mu,\gamma)\in\Theta\times\mathcal{M}_{GT}\times\Gamma_G}{\text{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\big(y_{it} - \overline{y}_i - (x_{it} - \overline{x}_i)'\theta - \mu_{g_i t}\big)^2, \tag{S.21}$$

[11]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 (S.20), we have simply added the number of group-specific time effects (i.e., GT), the number of common parameters (K), and the number of group membership variables g_i (i.e., N). Below we report simulation results using (S.20), as well as using an alternative choice with a steeper penalty.

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

$$\widehat{\sigma}^2 = \frac{1}{NT - G_{\max}T - N - K}\sum_{i=1}^{N}\sum_{t=1}^{T}\big(y_{it} - x_{it}'\widehat{\theta} - \widehat{\alpha}_{\widehat{g}_i t}\big)^2.$$

and

$$\left(\widetilde{\theta}^{FE}, \widetilde{\mu}^{FE}\right) = \underset{(\theta,\mu)\in\Theta\times\mathcal{M}_{GT}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \overline{y}_i - (x_{it} - \overline{x}_i)'\theta - \mu_{g_i^0 t}\right)^2, \tag{S.22}$$

where

$$\mathcal{M}_{GT} = \big\{\mu \in \mathbb{R}^{GT},\ \text{for some } \alpha \in \mathcal{A}^{GT}\ \mu_{gt} = \alpha_{gt} - \overline{\alpha}_g \text{ for all } (g,t) \in \{1,\dots,G\}\times\{1,\dots,T\}\big\}.$$

We denote $\widehat{\gamma}^{FE} = \{\widehat{g}_1^{FE}, \dots, \widehat{g}_N^{FE}\}$.

Consider the following assumptions.

ASSUMPTION S.1:

(a) *For all* $(g,\widetilde{g}) \in \{1,\dots,G\}^2$ *such that* $g \neq \widetilde{g}$: $\operatorname{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\alpha_{gt}^0 - \overline{\alpha}_g^0 - \alpha_{\widetilde{g}t}^0 + \overline{\alpha}_{\widetilde{g}}^0)^2 = c_{g,\widetilde{g}}^{FE} > 0$.

(b) *There exists a constant* $M^* > 0$ *such that, as* N, T *tend to infinity,*

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

Assumption S.1(a) is a group separation condition. Assumption S.1(b) is related to, but weaker than, Assumption 2(e). We have the following two results.

PROPOSITION S.3—Unit-Specific Heterogeneity: *Suppose that Assumptions* 1(a)–1(c) *hold, and that Assumptions* 1(d)–1(g) *hold with* x_{it} *and* v_{it} *replaced by* $x_{it} - \overline{x}_i$ *and* $v_{it} - \overline{v}_i$, *respectively. Suppose also that Assumptions* 2(a) *and* 2(c)–2(d) *hold, and that Assumption* S.1 *holds. Then*

$$\Pr\Big(\sup_{i\in\{1,\dots,N\}} \left|\widehat{g}_i^{FE} - g_i^0\right| > 0\Big) = o(1) + o\left(NT^{-\delta}\right), \tag{S.23}$$

and

$$\widehat{\theta}^{FE} = \widetilde{\theta}^{FE} + o_p\left(T^{-\delta}\right), \quad \textit{and} \tag{S.24}$$

$$\widehat{\mu}_{gt}^{FE} = \widetilde{\mu}_{gt}^{FE} + o_p\left(T^{-\delta}\right) \quad \textit{for all } g,t. \tag{S.25}$$

For the proof see Appendix S.A.5.

COROLLARY S.1—Unit-Specific Heterogeneity: *Suppose that the conditions of Proposition* S.3 *are satisfied. Suppose also that Assumption* 3 *holds, with* x_{it} *and* v_{it} *replaced by* $x_{it} - \overline{x}_i$ *and* $v_{it} - \overline{v}_i$, *respectively. Then, as* N *and* T *tend to infinity such that* $N/T^{\nu} \to 0$ *for some* $\nu > 0$, *we have*

$$\sqrt{NT}\left(\widehat{\theta}^{FE} - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \left[\Sigma_\theta^{FE}\right]^{-1}\Omega_\theta^{FE}\left[\Sigma_\theta^{FE}\right]^{-1}\right), \tag{S.26}$$

and, for all (g,t),

$$\sqrt{N}\big(\widehat{\mu}_{gt}^{FE} - \big(\alpha_{gt}^0 - \overline{\alpha}_g^0\big)\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}^{FE}}{\pi_g^2}\right), \tag{S.27}$$

where[13]

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

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

$$\omega_{gt}^{FE} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}\big(\mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_j^0 = g\}(v_{it} - \overline{v}_i)(v_{jt} - \overline{v}_j)\big).$$

PROOF: Essentially identical to the proof of Corollary 1. *Q.E.D.*

Note that the conditions of Proposition S.3 do not rule out the presence of lagged dependent variables in x_{it}. For example, if y_{it} follows a stable autoregressive process with i.i.d. innovations v_{it} in both dimensions, and group-time effects independent of v_{it}, it can be shown along the lines of Alvarez and Arellano (2003) that[14]

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathbb{E}\big[(v_{it} - \overline{v}_i)(v_{is} - \overline{v}_i)(y_{i,t-1} - \overline{y}_{i,-1})(y_{i,s-1} - \overline{y}_{i,-1})\big] \\
&\quad = T\mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}(v_{it} - \overline{v}_i)(y_{i,t-1} - \overline{y}_{i,-1})\right)^2\right] \\
&\quad = O(1),
\end{aligned}$$

so that Assumption 1(d) holds with x_{it} and v_{it} replaced by $x_{it} - \overline{x}_i$ and $v_{it} - \overline{v}_i$. Moreover, Assumption S.1(b) holds in the presence of lagged outcomes, under conditions that we provide in Appendix S.A.5.

In contrast, the conditions needed to apply Corollary S.1 rule out the presence of a lagged outcome in x_{it}. Indeed, the counterpart to Assumption 3(a) is

$$\mathbb{E}\big[(x_{jt} - \overline{x}_j)(v_{it} - \overline{v}_i)\big] = 0 \quad \text{for all } i, j, t.$$

[13]We denote $\overline{x}_g = \frac{\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\}\overline{x}_i}{\sum_{i=1}^N \mathbf{1}\{g_i^0 = g\}}$, for all $g \in \{1, \ldots, G\}$.

[14]Here we denote: $\overline{y}_{i,-1} = \frac{1}{T}\sum_{t=1}^{T} y_{i,t-1}$.

This condition holds when covariates are strictly exogenous, but fails to hold in models with lagged outcomes where $\mathbb{E}[(y_{i,t-1} - \overline{y}_{i,-1})(v_{it} - \overline{v}_i)] = O(1/T)$ is not zero in general.

Instrumental Variables

When x_{it} are not strictly exogenous (yet the conditions of Proposition S.3 hold), one possibility is to use an instrumental variables strategy in the first-differenced equation:

$$\Delta y_{it} = \Delta x_{it}'\theta^0 + \Delta \alpha^0_{g_i^0 t} + \Delta v_{it}, \tag{S.28}$$

where $\Delta w_{it} = w_{it} - w_{i,t-1}$. This approach relies on the availability of a vector of instruments z_{it} such that $\mathbb{E}(z_{jt}\Delta v_{it}) = 0$ for all i, j, t. For example, when $x_{it} = (y_{i,t-1}, \widetilde{x}_{it}')'$ contain a lagged outcome and a vector of strictly exogenous covariates, and when v_{it} are independent across units and over time, we can take $z_{it} = (y_{i,t-2}, \Delta\widetilde{x}_{it}')'$, in analogy with IV techniques commonly used in linear panel data models (Anderson and Hsiao (1982)).

If population group membership indicators were known, one could consider the following IV estimator of θ^0, which uses z_{it} and interactions of group and time dummies as instruments:[15]

$$\widetilde{\theta}^{IV} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}(\Delta x_{it} - \overline{\Delta x}_{g_i^0 t})'\right]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \overline{\Delta y}_{g_i^0 t}). \tag{S.29}$$

Under standard conditions, $\sqrt{NT}(\widetilde{\theta}^{IV} - \theta^0)$ is asymptotically distributed as $\mathcal{N}(0, V_{IV})$ as N and T tend to infinity.

A feasible counterpart to $\widetilde{\theta}^{IV}$ is given by

$$\widehat{\theta}^{IV} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}(\Delta x_{it} - \overline{\Delta x}_{\widehat{g}_i^{FE} t})'\right]^{-1} \sum_{i=1}^{N}\sum_{t=1}^{T} z_{it}(\Delta y_{it} - \overline{\Delta y}_{\widehat{g}_i^{FE} t}), \tag{S.30}$$

where the GFE estimates $\widehat{g}_i^{FE}$ are given by (S.21). Using (S.23), one can show that, as N and T tend to infinity such that N/T^ν tends to zero for some $\nu > 0$, $\sqrt{NT}(\widehat{\theta}^{IV} - \widetilde{\theta}^{IV}) = o_p(1)$.[16] As a result, $\sqrt{NT}(\widehat{\theta}^{IV} - \theta^0) \xrightarrow{d} \mathcal{N}(0, V_{IV})$. An analogous result holds for the IV estimates of $\Delta\alpha^0_{gt}$.

[15] For simplicity, we focus on the just-identified case where z_{it} and x_{it} have the same dimension.

[16] Indeed, for all $\varepsilon > 0$, if $N/T^\nu \to 0$ for some $\nu > 0$,

$$\Pr\big[\big|\sqrt{NT}\big(\widehat{\theta}^{IV} - \widetilde{\theta}^{IV}\big)\big| > \varepsilon\big] \leq \Pr\Big(\sup_{i\in\{1,\ldots,N\}} \big|\widehat{g}_i^{FE} - g_i^0\big| > 0\Big) = o(1).$$

S.4.2. *Extension 2: Heterogeneous Coefficients*

Consider the following extension of model (12) with heterogeneous coefficients:

$$y_{it} = x_{it}'\theta^0_{g_i^0} + \alpha^0_{g_i^0 t} + v_{it}. \tag{S.31}$$

In this model, the GFE estimator is defined as[17]

$$\left(\widehat{\theta}^{HC}, \widehat{\alpha}^{HC}, \widehat{\gamma}^{HC}\right) = \underset{(\theta,\alpha,\gamma)\in\Theta^G\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta_{g_i} - \alpha_{g_i t}\right)^2. \tag{S.32}$$

We denote $\widehat{\gamma}^{HC} = \{\widehat{g}_1^{HC}, \ldots, \widehat{g}_N^{HC}\}$.

Let us also define the infeasible counterpart to $(\widehat{\theta}^{HC}, \widehat{\alpha}^{HC})$ as

$$\left(\widetilde{\theta}^{HC}, \widetilde{\alpha}^{HC}\right) = \underset{(\theta,\alpha)\in\Theta^G\times\mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta_{g_i^0} - \alpha_{g_i^0 t}\right)^2. \tag{S.33}$$

Consider the following set of assumptions.

ASSUMPTION S.2:

(a) *There exists a* $\widehat{\rho}^{HC} \xrightarrow{p} \rho^{HC} > 0$ *such that, for all* g, $\min_{\gamma\in\Gamma_G} \max_{\widetilde{g}\in\{1,\ldots,G\}} \widehat{\rho}(\gamma, g, \widetilde{g}) \geq \widehat{\rho}^{HC}$, *where* $\widehat{\rho}(\gamma, g, \widetilde{g})$ *is the minimum eigenvalue of the following* $(K+T)\times(K+T)$ *matrix* (*with* $K = \dim x_{it}$):

$$M(\gamma, g, \widetilde{g}) \equiv \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_i = \widetilde{g}\} \times \begin{pmatrix} \frac{1}{T}\sum_{t=1}^{T} x_{it}x_{it}' & \frac{1}{\sqrt{T}}x_{i1} & \frac{1}{\sqrt{T}}x_{i2} & \cdots & \frac{1}{\sqrt{T}}x_{iT} \\ \frac{1}{\sqrt{T}}x_{i1}' & 1 & 0 & \cdots & 0 \\ \frac{1}{\sqrt{T}}x_{i2}' & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{1}{\sqrt{T}}x_{iT}' & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

[17]Extensions of our algorithms can be used to compute the GFE estimator in (S.32). See, for example, the literature on "clusterwise regression" in operations research (Späth (1979), Caporossi and Hansen (2005), and more recently, Lin and Ng (2012)). The results of the heterogeneous coefficients model that we report in the empirical section below are based on a counterpart to Algorithm 1.

(b) *For all* $g \neq \widetilde{g}$, *there exists a* $c_{g,\widetilde{g}}^{HC} > 0$ *such that* $\operatorname{plim}_{N,T\to\infty} \frac{1}{N}\sum_{i=1}^{N} D_{g\widetilde{g}i}^{0} \geq c_{g,\widetilde{g}}^{HC}$ *and, for all* $i \in \{1,\dots,N\}$, $\operatorname{plim}_{T\to\infty} D_{g\widetilde{g}i}^{0} \geq c_{g,\widetilde{g}}^{HC}$, *where* $D_{g\widetilde{g}i}^{0} = \frac{1}{T}\sum_{t=1}^{T}(x_{it}'(\theta_{g}^{0} - \theta_{\widetilde{g}}^{0}) + \alpha_{gt}^{0} - \alpha_{\widetilde{g}t}^{0})^{2}$.

(c) *There exists a constant* $M^{*} > 0$ *such that*

$$\sup_{i\in\{1,\dots,N\}} \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^{2} \geq M^{*}\right) = o\big(T^{-\delta}\big) \quad \textit{for all } \delta > 0.$$

(d) *For all constants* $c > 0$,

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

Assumption S.2(a) is a relevance condition related to Assumption 1(g) in the paper. In particular, in analogy with the baseline model, this condition fails when, for some g and γ, the matrix

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{1}\{g_{i}^{0} = g\}\mathbf{1}\{g_{i} = \widetilde{g}\}(x_{it} - \overline{x}_{g\wedge\widetilde{g},t})(x_{it} - \overline{x}_{g\wedge\widetilde{g},t})'$$

is singular for all $\widetilde{g}$.

To provide some intuition for Assumption S.2(a), let us consider the case where x_{it} are scalar standard normal, i.i.d. in both dimensions. For a given partition γ, and given $(g,\widetilde{g})$, $M(\gamma,g,\widetilde{g})$ takes the form

$$M(\gamma,g,\widetilde{g}) = \begin{pmatrix} \widehat{a} & \dfrac{1}{\sqrt{T}}\widehat{b}' \\ \dfrac{1}{\sqrt{T}}\widehat{b} & \dfrac{N(g,\widetilde{g})}{N} I_{T} \end{pmatrix},$$

$$\widehat{a} \sim \frac{\chi^{2}_{N(g,\widetilde{g})T}}{NT}, \quad \text{and} \quad \widehat{b}_{t} \sim \frac{\mathcal{N}\big(0, N(g,\widetilde{g})\big)}{N} \quad \text{for all } t \in \{1,\dots,T\},$$

where I_{T} denotes the $T \times T$ identity matrix, and $N(g,\widetilde{g}) \equiv \sum_{i=1}^{N}\mathbf{1}\{g_{i}^{0} = g\} \times \mathbf{1}\{g_{i} = \widetilde{g}\}$. Moreover, by Assumption 2(a), $\max_{\widetilde{g}\in\{1,\dots,G\}} \frac{N(g,\widetilde{g})}{N} \geq \frac{\pi_{g}}{2G}$ with probability approaching 1. Hence, for a suitable choice of $\widetilde{g}$, $\mathbb{E}[M(\gamma,g,\widetilde{g})]$ is asymptotically bounded from below by a positive constant times I_{T+1}. In Appendix S.A.6, we formally show that Assumption S.2(a) holds in this case. This requires taking into account the minimization with respect to γ (which introduces a G^{N} probability factor), and showing that the rates of convergence of $\widehat{a}$ and $\frac{1}{\sqrt{T}}\widehat{b}$ are sufficiently fast to dominate in the limit.

Assumption S.2(b) is a group separation condition. As in the baseline model, this condition is instrumental to derive asymptotic equivalence. A difference with Assumptions 2(b) and S.1(a) is that, in this case, the condition depends

on the data $(x_{i1}, \ldots, x_{iT})$. Intuitively, it is satisfied if, for all i and $\widetilde{g} \neq g$, $\{x_{it}\}_t$ and $\{\alpha_{gt}^0 - \alpha_{\widetilde{g}t}^0\}_t$ are not collinear. For example, if $x_{it} = x_i$ are time-invariant, then Assumption S.1(a) implies Assumption S.2(b). In practice, however, Assumption S.2(b) might fail to hold for some units, exactly or approximately. One would then expect group classification to be inaccurate for those units. Providing methods to test for group separation, and to achieve valid inference on the model's parameters when it fails, are interesting questions for future work.

Assumption S.2(c) is a slightly more restrictive version of Assumption 2(e). Lastly, Assumption S.2(d) imposes a condition on the tail properties of $\frac{1}{T}\sum_{t=1}^{T} v_{it}x_{it}$. It will hold if, similarly as v_{it}, $v_{it}x_{it}$ satisfy mixing and tail conditions of the form given in Assumption 2 in the paper.[18]

We then have the following result.

PROPOSITION S.4—Heterogeneous Coefficients: *Suppose that Assumptions* 1(a)–1(f), *Assumptions* 2(a) *and* 2(c)–2(d), *and Assumption* S.2 *hold*. *Then*, *as* N, T *tend to infinity*,

$$\Pr\Big(\sup_{i\in\{1,\ldots,N\}} \big|\widehat{g}_i^{HC} - g_i^0\big| > 0\Big) = o(1) + o\big(NT^{-\delta}\big), \tag{S.34}$$

and

$$\widehat{\theta}_g^{HC} = \widetilde{\theta}_g^{HC} + o_p\big(T^{-\delta}\big) \quad \textit{for all } g, \quad \textit{and} \tag{S.35}$$

$$\widehat{\alpha}_{gt}^{HC} = \widetilde{\alpha}_{gt}^{HC} + o_p\big(T^{-\delta}\big) \quad \textit{for all } g, t. \tag{S.36}$$

For the proof see Appendix S.A.6.

We also have a result analogous to Corollary 1, which we give without proof.

ASSUMPTION S.3:

(a) *For all* $i, j, t,$ *and* g, $\mathbb{E}(\mathbf{1}\{g_i^0 = g\}x_{jt}v_{it}) = 0$.

(b) *For all* g, *there exist positive definite matrices* $\Sigma_{\theta g}^{HC}$ *and* $\Omega_{\theta g}^{HC}$ *such that*

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

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

[18]Lagged outcomes may also be strongly mixing under additional conditions. For example, the conditions in Chanda (1974) for linear stochastic processes to be strongly mixing involve restrictions on the characteristic functions of innovations. However, here we do not provide primitive conditions for Assumption S.2(d) in models with lagged outcomes.

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

COROLLARY S.2—Heterogeneous Coefficients: *Suppose that the conditions of Proposition S.4, Assumptions* 3(d)–3(e), *and Assumption S.3 hold, and let N and T tend to infinity such that, for some $\nu > 0$, $N/T^{\nu} \to 0$. Then we have, for all g,*

$$\sqrt{NT}\big(\widehat{\theta}_g^{HC} - \theta_g^0\big) \xrightarrow{d} \mathcal{N}\big(0, \big[\Sigma_{\theta g}^{HC}\big]^{-1}\Omega_{\theta g}^{HC}\big[\Sigma_{\theta g}^{HC}\big]^{-1}\big), \tag{S.37}$$

and, for all (g, t),

$$\sqrt{N}\big(\widehat{\alpha}_{gt}^{HC} - \alpha_{gt}^0\big) \xrightarrow{d} \mathcal{N}\left(0, \frac{\omega_{gt}}{\pi_g^2}\right). \tag{S.38}$$

S.5. COMPLEMENTS TO THE MAIN ANALYSIS

In this section, we study four issues in turn: the link between GFE and finite mixtures, how to incorporate prior information in GFE estimation, how to fit a model to the estimated groups, and GFE estimation in unbalanced panels.

S.5.1. *Connection to Finite Mixture Models*

Here we show that the grouped fixed-effects estimator in model (1) can be interpreted as the maximizer of the pseudo-likelihood of a mixture-of-normals model, where the mixing probabilities are individual-specific and unrestricted. This contrasts with standard finite mixture modeling (McLachlan and Peel (2000)), which typically specifies the group probabilities $\pi_g(x_i)$ as functions of the covariates. In comparison, in the grouped fixed-effects approach, the group probabilities $\pi_{ig} = \pi_g(i)$ are unrestricted functions of the individual dummies.

To state the equivalence result, let $\sigma > 0$ be a scaling parameter. Then, it is easy to see that the GFE estimator $(\widehat{\theta}, \widehat{\alpha})$ given by equation (2) in the paper satisfies

$$\begin{aligned}(\widehat{\theta}, \widehat{\alpha}) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmax}}\Bigg[\max_{\pi_1,\ldots,\pi_N}\sum_{i=1}^{N}\ln\Bigg(\sum_{g=1}^{G}\pi_{ig}\frac{1}{\big(2\pi\sigma^2\big)^{T/2}} \\ \times\exp\Bigg(-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\big(y_{it} - x_{it}'\theta - \alpha_{gt}\big)^2\Bigg)\Bigg)\Bigg],\end{aligned} \tag{S.39}$$

where the maximum is taken over all probability vectors $\pi_i = (\pi_{i1}, \ldots, \pi_{iG})$ in the unit simplex of $\mathbb{R}^G$. Result (S.39) comes from the fact that the individual-specific π_i are unrestricted. Specifically, given (θ, α) values, the maximum is

achieved at

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

that is,

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

Note that (S.39) holds for any choice of σ.

S.5.2. *Adding Prior Information*

Modeling Time Patterns

A simple extension of the benchmark grouped fixed-effects model is to impose linear constraints on the group-specific time effects α_{gt}. For example, one may specify $\alpha_{gt} = \sum_{r=1}^{R} b_{gr}\psi_r(t)$, where $\psi_1, \ldots, \psi_R$ are known functions, and where b_{gr} are scalar parameters to be estimated. Linear constraints are easy to embed within the computational and statistical framework of model (1), and allow to model a wide variety of patterns of unobserved heterogeneity.

In the empirical application below, we show estimates of a model with two different layers of heterogeneity (time-varying and time-invariant) that takes the following form:

$$y_{it} = x_{it}'\theta + \alpha_{g_{i1}t} + \eta_{g_{i1},g_{i2}} + v_{it}, \tag{S.40}$$

where $(g_{i1}, g_{i2}) \in \{1, \ldots, G_1\} \times \{1, \ldots, G_2\}$ indicates joint group membership. Model (S.40) may be interpreted as a restricted version of model (1) with $G = G_1 \times G_2$ groups, and with linear constraints on the group-specific time effects. Indeed, letting $\mu_{g_1g_2t} = \alpha_{g_1t} + \eta_{g_1,g_2}$, it is easy to see that the following $G_1(G_2-1)(T-1)$ linear constraints are satisfied:

$$\mu_{g_1g_2t} - \frac{1}{T}\sum_{s=1}^{T}\mu_{g_1g_2s} - \frac{1}{G_2}\sum_{h=1}^{G_2}\mu_{g_1ht} + \frac{1}{G_2T}\sum_{h=1}^{G_2}\sum_{s=1}^{T}\mu_{g_1hs} = 0 \quad \text{for all } (g_1, g_2, t).$$

Prior Information on Group Membership

In certain applications, researchers may want to incorporate prior information on the structure of unobserved heterogeneity. For example, in a cross-country application, one could think that countries in the same continent are

more likely to belong to the same group. In such situations, one possibility is to impose the grouped structure on the data by assumption, for example, by allowing for continent fixed-effects possibly interacted with time effects. Another approach is to use our grouped fixed-effects estimator, which leaves the groups unrestricted and recovers them endogenously. An intermediate possibility is to combine a priori information on group membership with data information, simply by adding a penalty term to the GFE objective.

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

$$\left(\widehat{\theta}^{(\pi)}, \widehat{\alpha}^{(\pi)}\right) = \underset{(\theta,\alpha)\in\Theta\times\mathcal{A}^{GT}}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{\widehat{g}_i^{(\pi_i)}(\theta,\alpha)t}\right)^2, \tag{S.41}$$

where the estimated group variables are now

$$\widehat{g}_i^{(\pi_i)}(\theta,\alpha) = \underset{g\in\{1,\ldots,G\}}{\operatorname{argmin}} \sum_{t=1}^{T}\left(y_{it} - x_{it}'\theta - \alpha_{gt}\right)^2 - C\ln\pi_{ig}, \tag{S.42}$$

and where $C > 0$ is a penalty term. The penalty specifies the respective weights attached to prior and data information in estimation.[19]

Note that computation of the penalized GFE estimator is very similar to that of the baseline GFE estimator. In addition, the penalized and unpenalized GFE estimators are asymptotically equivalent under the conditions given in Section 3, provided prior information is nondogmatic in the following sense:

ASSUMPTION S.4—Prior Probabilities: *For some* $\varepsilon > 0$,

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

We have the following result.

COROLLARY S.3—Penalized GFE: *Let the assumptions of Corollary* 1 *hold, and let* $\pi = \{\pi_{ig}\}$ *be a set of prior probabilities that satisfies Assumption* S.4. *Then, as* N *and* T *tend to infinity such that* $N/T^{\nu} \to 0$ *for some* $\nu > 0$,

$$\sqrt{NT}\left(\widehat{\theta}^{(\pi)} - \theta^0\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1}\right). \tag{S.43}$$

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

PROOF: The proof closely follows that of Theorem 2 and Corollary 1. A difference appears in the proof of Lemma B4. Let us define the following quantity:

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

The proof consists in bounding $Z_{ig}^{(\pi_i)}(\theta,\alpha)$ instead of bounding $Z_{ig}(\theta,\alpha)$. The only difference is that A_T has the following extra term: $|-C\ln\pi_{ig} + C\ln\pi_{i\tilde{g}}|$, which is bounded by $C\ln(\frac{1-\varepsilon}{\varepsilon})$ by Assumption S.4. *Q.E.D.*

In standard fixed-effects models, adding prior information on the individual effects adds generally an $O(1/T)$ term to the small-T bias of the estimator (Arellano and Bonhomme (2009)). In contrast, Corollary S.3 shows that, in models where unobserved heterogeneity is discrete and the number of groups is fixed, and under the conditions of Theorem 2, adding nondogmatic prior information has no effect on the first-order asymptotic distribution of the estimator as N and T tend to infinity and N/T^{ν} tends to zero. It is worth noting, however, that prior information will impact the higher-order and finite-sample properties of the GFE estimator.

S.5.3. *Fitting a Probability Model to the Estimated Groups*

Suppose one wants to fit a parametric model (e.g., an ordered probit or a multinomial logit model), indexed by a parameter vector ξ, to the estimated groups:

$$
\widehat{\xi} = \operatorname*{argmax}_{\xi} \sum_{i=1}^{N} \sum_{g=1}^{G} \mathbf{1}\{\widehat{g}_i = g\} \ln\bigl(p_g(x_i;\xi)\bigr),
$$

where $p_g(x;\xi)$ are the parametrically specified group probabilities. For example, in the empirical application below, we will use a multinomial logit model to link the estimated groups to country-specific determinants. It is easy to see that, in a large-N, T perspective and under similar conditions as in Theorem 2, $\widehat{\xi}$ will be asymptotically equivalent to the following infeasible maximum likelihood estimator:

$$
\widetilde{\xi} = \operatorname*{argmax}_{\xi} \sum_{i=1}^{N} \sum_{g=1}^{G} \mathbf{1}\{g_i^0 = g\} \ln\bigl(p_g(x_i;\xi)\bigr).
$$

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

S.5.4. *Grouped Fixed-Effects in Unbalanced Panels*

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

The GFE estimator is then

$$(\widehat{\theta}, \widehat{\alpha}, \widehat{\gamma}) = \underset{(\theta,\alpha,\gamma)\in\Theta\times\mathcal{A}^{GT}\times\Gamma_G}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} d_{it}\big(y_{it} - x_{it}'\theta - \alpha_{g_i t}\big)^2. \tag{S.44}$$

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

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

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

S.6. SIMULATION EXERCISES

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

S.6.1. *Design and Main Results*

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

TABLE S.III

BIAS OF THE GFE ESTIMATOR[a]

	θ_1 (Coeff. $y_{i,t-1}$)		θ_2 (Coeff. $\widetilde{x}_{it}$)		$\frac{\theta_2}{1-\theta_1}$		
	True	GFE	True	GFE	True	GFE	Misclassified
$G=3$	0.407	0.391	0.089	0.099	0.151	0.163	9.50%
$G=5$	0.255	0.262	0.079	0.086	0.107	0.117	9.68%
$G=10$	0.277	0.286	0.075	0.078	0.104	0.110	44.73%

[a]Model (12) with G groups. The columns labeled "GFE" refer to the mean of GFE parameter estimates across 1,000 simulations. Algorithm 2—with parameters (5; 10; 5)—was used for computation. The last column shows the average of the misclassification frequency ($\widehat{g}_i \neq g_i^0$) across simulations. Errors are i.i.d. normal.

across units and time periods, with variance equal to the mean of squared GFE residuals.

We start by showing the mean of the GFE estimator across 1,000 Monte Carlo simulations in Table S.III.[20] We show the results for the two coefficients (θ_1 and θ_2, respectively), as well as for the "long-run" coefficient of $\widetilde{x}_{it}$ (i.e., $\theta_2/(1-\theta_1)$). Here the number of groups used in estimation (G) is the same as the true number of groups (G^0). Biases appear moderate despite the short length of the panel, at most 10% in relative terms. The last column in Table S.III shows the average misclassification frequency across simulations.[21] When $G=3$ or 5, units are well classified in approximately 90% of cases. When $G=10$, however, the frequency of correct classification drops to 55%. Nonetheless, the bias of the GFE estimator remains rather low. This suggests that the GFE estimator of common parameters may behave well in situations when G is not small relative to the sample size.

We next turn to inference. The top panel in Table S.IV reports the standard deviation of the GFE estimator of θ across Monte Carlo simulations, together with the medians across simulations of three different standard errors estimates: the (square root of the) clustered variance formula (S.3), estimates based on Pollard's (1982) fixed-T formula, and estimates based on the bootstrap (computed by resampling unit-specific sequences of observations with replacement).[22] All variance formulas are robust to the presence of serial correlation, but rely on the assumption that observations are independent across units. The results show that the clustered formula based on a large-T approximation systematically underpredicts the variability of the GFE estimator. This

[20]Medians across simulations are almost identical to the means (not reported).

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

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

TABLE S.IV

INFERENCE FOR THE GFE ESTIMATOR[a]

Standard Errors

	θ_1 (Coeff. $y_{i,t-1}$)				θ_2 (Coeff. $\widetilde{x}_{it}$)				$\frac{\theta_2}{1-\theta_1}$			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$G=3$	0.035	0.051	0.068	0.043	0.0093	0.0132	0.0156	0.0137	0.013	0.022	0.030	0.021
$G=5$	0.037	0.068	0.097	0.058	0.0088	0.0135	0.0160	0.0112	0.011	0.022	0.035	0.022
$G=10$	0.037	0.048	0.091	0.059	0.0074	0.0095	0.0156	0.0103	0.009	0.012	0.026	0.015

Coverage (Nominal Level 5%)

	θ_1 (Coeff. $y_{i,t-1}$)			θ_2 (Coeff. $\widetilde{x}_{it}$)			$\frac{\theta_2}{1-\theta_1}$		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
$G=3$	0.847	0.965	0.970	0.723	0.883	0.914	0.693	0.917	0.900
$G=5$	0.848	0.961	0.973	0.788	0.932	0.943	0.710	0.936	0.928
$G=10$	0.798	0.902	0.992	0.841	0.912	0.996	0.783	0.904	0.986

[a]Model (12) with G groups. Median standard errors across 1,000 simulations (top panel) and empirical nonrejection probabilities (bottom panel, nominal size 5%). Column (1) reports results based on the large-T clustered variance formula (S.3), (2) reports estimates based on Pollard's (1982) fixed-T formula, and (3) shows results based on the bootstrap (100 replications, Algorithm 1 with 1,000 starting values). Column (4) in the top panel shows Monte Carlo standard deviations across simulations. Errors are i.i.d. normal.

shows that group misclassification may have a sizable effect on inference in small samples. In contrast, the two consistent estimates of the fixed-T variance are larger, and more in line with the finite-sample dispersion. Moreover, the bottom panel in the table shows that these two methods provide approximately correct coverage for the true parameter θ^0, while estimates based on the large-T approximation tend to lead to overrejection.

S.6.2. *Additional Results*

Here we show the results of several additional exercises.

Comparison With Interactive Fixed-Effects

We first consider an alternative estimator, the interactive fixed-effects estimator of Bai (2009) with three factors, when the DGP follows the GFE model (1) with $G=3$ groups. Note that model (1) can be written as

$$y_{it} = x_{it}'\theta + \lambda_i' f_t + v_{it}, \tag{S.45}$$

where f_t and λ_i are $G \times 1$, and $\lambda_{ig}^0 = \mathbf{1}\{g_i^0 = g\}$ and $f_{tg}^0 = \alpha_{gt}^0$ for all g, i, t.

Table S.V shows the results of 1,000 Monte Carlo replications, and compares the interactive fixed-effects estimator to the GFE estimator with $G=3$ groups.

TABLE S.V

COMPARISON WITH INTERACTIVE FIXED-EFFECTS ($G = 3$)[a]

		GFE		IFE	
	True	Mean	Std.	Mean	Std.
θ_1 (coeff. $y_{i,t-1}$)	0.407	0.391	0.043	−0.329	0.040
θ_2 (coeff. $\widetilde{x}_{it}$)	0.089	0.099	0.014	0.146	0.035
$\frac{\theta_2}{1-\theta_1}$	0.151	0.163	0.021	0.110	0.026
		Mean	Median	Mean	Median
$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{c}_{it} - \alpha^0_{g_i^0 t})^2$	–	0.023	0.014	0.164	0.099

[a]Model (12) with $G = 3$ groups. GFE is the grouped fixed-effects estimator, IFE is the interactive fixed-effects estimator (computed using true parameter values as starting conditions). $\widehat{c}_{it} = \widehat{\alpha}_{\widehat{g}_i t}$ in GFE, and $\widehat{c}_{it} = \widehat{\lambda}_i' \widehat{f}_t$ in IFE. Means, medians, and standard deviations across 1,000 simulations. Errors are i.i.d. normal.

Although the interactive fixed-effects estimator is consistent as N and T tend to infinity, the first three rows of the table show that it suffers from a very substantial finite sample bias, much larger than the bias of the GFE estimator on this (relatively small) sample. Specifically, the mean of the autoregressive parameter and the coefficient of $\widetilde{x}_{it}$ are −0.329 and 0.146, whereas the true values are 0.407 and 0.089, respectively. This result is consistent with the theoretical properties of GFE and interactive fixed-effects: while the former is unbiased as $N/T^\nu \to \infty$ for some $\nu > 0$, the latter generally suffers from a $O(1/T)$ bias even as N/T tends to a constant.[23]

The last row in Table S.V shows the mean and median across simulations of the following average of squared errors of the estimated components of unobserved heterogeneity:

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\widehat{c}_{it} - \alpha^0_{g_i^0 t}\right)^2,$$

where $\widehat{c}_{it} = \widehat{\alpha}_{\widehat{g}_i t}$ in GFE, and $\widehat{c}_{it} = \widehat{\lambda}_i' \widehat{f}_t$ in interactive fixed-effects. The results show large differences between the two estimators. Considering the median across simulations, the difference $|\widehat{\alpha}_{\widehat{g}_i t} - \alpha^0_{g_i^0 t}|$ is $\sqrt{0.014} \approx 0.12$ on average. In contrast, $|\widehat{\lambda}_i' \widehat{f}_t - \alpha^0_{g_i^0 t}|$ is $\sqrt{0.099} \approx 0.32$ on average. Hence, in this design, interactive fixed-effects yields imprecise estimates of the components of unobserved

[23]Bai (2009) discussed bias reduction in interactive fixed-effects models with strictly exogenous regressors. Moon and Weidner (2010b) provided truncation-based bias reduction formulas in models with predetermined regressors. Note that, in contrast with interactive fixed-effects, under the conditions of Corollary 1, the small-T bias of the GFE estimator vanishes at a faster-than-polynomial rate, even in the presence of lagged outcomes.

heterogeneity when compared to GFE. Moreover, in Section 3.3 of the paper we showed that the theoretical rate of convergence of $\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(\widehat{c}_{it}-\alpha^0_{g_i^0 t})^2$ is $O_p(1/N)$ in GFE, compared to $O_p(1/\min(N,T))=O_p(1/T)$ in interactive fixed-effects. The results reported at the bottom of Table S.V are in line with these theoretical rates, as 0.32/0.12 and $\sqrt{N/T}$ are of a similar order of magnitude. Overall, this comparison suggests that the more parsimonious GFE estimator may outperform interactive fixed-effects in relatively short panels when the data have a grouped structure.

Group-Specific Time Effects

Turning next to group-specific time effects, Figure S.1 shows the pointwise means of $\widehat{\alpha}_{gt}$ across 1,000 simulations. Both when $G=3$ and when $G=5$, all time profiles are shifted downwards relative to the true ones. Nevertheless, the overall patterns of heterogeneity are well reproduced. In fact, we checked that the group-specific means of y_{it} and x_{it} are almost unbiased (not reported).[24]

Nonnormal Design

The main simulation results are based on a design with i.i.d. normal errors, which might seem too favorable given that the asymptotic behavior of the GFE estimator crucially depends on tail and dependence properties of errors. To address this concern, we report results using a different DGP, in which errors are resampled (with replacement) from the unit-specific vectors of GFE residu-

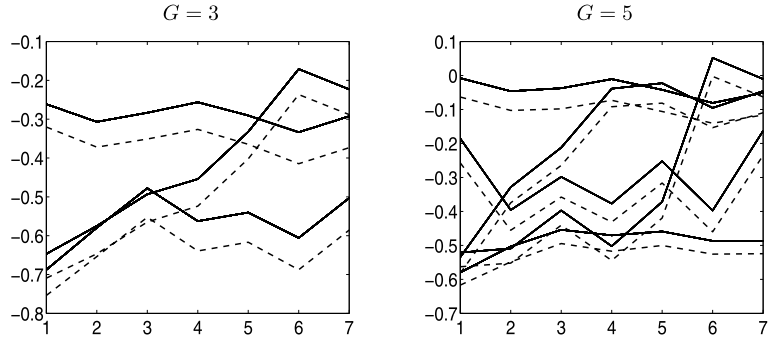

FIGURE S.1.—Monte Carlo bias of group-specific time effects. *Note*: Model (12) with G groups. Solid line shows the true values α^0_{gt}, dashed lines show the mean of $\widehat{\alpha}_{gt}$ across 1,000 simulations with i.i.d. normal errors. x-axis shows time $t\in\{1,\ldots,7\}$.

[24]We also computed the finite-sample variances of the group-specific time effects, and compared them with the clustered estimator (S.2) based on a large-N,T approximation. As in Table S.IV, the results show some sizable differences between the two.

TABLE S.VI

BIAS OF THE GFE ESTIMATOR (ALTERNATIVE DGP)[a]

	θ_1 (Coeff. $y_{i,t-1}$)		θ_2 (Coeff. $\widetilde{x}_{it}$)		$\frac{\theta_2}{1-\theta_1}$		
	True	GFE	True	GFE	True	GFE	Misclassified
$G=3$	0.407	0.381	0.089	0.099	0.151	0.163	9.86%
$G=5$	0.255	0.314	0.079	0.082	0.107	0.125	13.50%
$G=10$	0.277	0.322	0.075	0.074	0.104	0.109	33.27%

[a]Model (12) with G groups. The columns labeled "GFE" refer to the mean of GFE parameter estimates across 1,000 simulations. Algorithm 2—with parameters $(5; 10; 5)$—was used for computation. The last column shows the average of the misclassification frequency $(\widehat{g}_i \neq g_i^0)$ across simulations. Unit-specific sequences of errors are drawn with replacement from the estimated GFE residuals.

als.[25] Note that, given the nature of the original data, these residuals exhibit serial correlation and are clearly not normally distributed. Tables S.VI and S.VII report the mean and the standard deviation and coverage of the GFE estimator for θ, respectively, across 1,000 simulations. Compared with the i.i.d. normal case, the results show slightly larger small-sample biases, and a stronger underestimation of the finite-sample variance when using the formula based on large-T approximation. At the same time, Pollard's fixed-T formula and the bootstrap yield more accurate inference.[26]

Estimated Number of Groups

Additionally, we check the performance of the BIC criterion (S.20) to estimate the number of groups. To do so, we count the number of times that BIC selects a given G, across 100 simulated data sets. The results reported in Table S.VIII suggest that the criterion performs reasonably well, even in cases where the true number of groups is relatively large ($G^0 = 10$).[27]

We also run simulations where the number of groups G used in estimation differs from the true number G^0. Figure S.2 shows that the mean and standard deviation of the GFE estimator of common parameters do not differ much when $G > G^0$ compared to when $G^0 = 3$, consistently with the discussion in Section S.3, although we observe some increase in the finite-sample dispersion of the estimator as G grows.

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

[26]The results for group-specific time effects are similar to those shown in Figure S.1 (not reported).

[27]We also tried the alternative choice $\widehat{\sigma}^2 \frac{G(T+N-G)}{NT} \ln(NT)$ for the penalty, instead of $\widehat{\sigma}^2 \frac{GT+N+K}{NT} \ln(NT)$ in equation (S.20). This corresponds to a common choice of penalty in factor models (e.g., Bai and Ng (2002)). We found that, in this case, BIC selected 1 group in all 100 simulations, when the truth was $G^0 = 3$. In comparison, Table S.VIII shows that our choice (S.20) yields better results on these data.

TABLE S.VII

INFERENCE FOR THE GFE ESTIMATOR (ALTERNATIVE DGP)[a]

Standard Errors

	θ_1 (Coeff. $y_{i,t-1}$)				θ_2 (Coeff. $\widetilde{x}_{it}$)				$\frac{\theta_2}{1-\theta_1}$			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$G=3$	0.050	0.068	0.146	0.118	0.0104	0.0129	0.0179	0.0162	0.011	0.018	0.041	0.028
$G=5$	0.042	0.074	0.137	0.125	0.0083	0.0108	0.0126	0.0103	0.010	0.018	0.041	0.033
$G=10$	0.038	0.050	0.092	0.064	0.0067	0.0082	0.0115	0.0086	0.008	0.011	0.021	0.013

Coverage (Nominal Level 5%)

	θ_1 (Coeff. $y_{i,t-1}$)			θ_2 (Coeff. $\widetilde{x}_{it}$)			$\frac{\theta_2}{1-\theta_1}$		
	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
$G=3$	0.637	0.792	0.983	0.733	0.835	0.994	0.715	0.906	0.911
$G=5$	0.689	0.837	0.875	0.887	0.956	0.986	0.685	0.855	0.883
$G=10$	0.701	0.862	0.935	0.859	0.929	0.989	0.821	0.934	0.986

[a]Model (12) with G groups. Median standard errors across 1,000 simulations (top panel) and empirical nonrejection probabilities (bottom panel, nominal size 5%). Column (1) is based on the large-T variance formula, (2) is based on Pollard's (1982) fixed-T formula, (3) is based on the bootstrap, and (4) in the top panel shows Monte Carlo standard deviations across simulations. Unit-specific sequences of errors are drawn with replacement from the estimated GFE residuals.

Lastly, Table S.IX shows the mean and standard deviation of the GFE estimator across 100 simulations, when the number of groups is estimated using BIC in every simulation. The results on common parameter estimates show small differences compared to the results obtained with known G^0 (see Tables S.III and S.IV). At the same time, misspecification of the number of groups has important consequences for inference on the group-specific time

TABLE S.VIII

CHOICE OF THE NUMBER OF GROUPS, BIC[a]

		$G^0=3$					
	$G=$	1	2	3	4	5	6
$\%(\widehat{G}=G)$		0	0	98	2	0	0
		$G^0=10$					
	$G=$	7	8	9	10	11	12
$\%(\widehat{G}=G)$		0	10	42	42	6	0

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

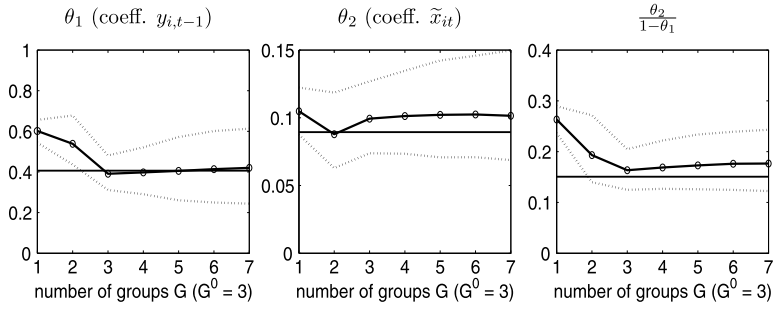

FIGURE S.2.—GFE based on G groups, when $G^0 = 3$. *Note*: See the notes to Table S.III. Model (12) with $G^0 = 3$ groups. GFE estimates are computed using G groups, where G is reported on the x-axis. Solid thick lines and dashed lines indicate the mean and 95% pointwise confidence bands, respectively, across 1,000 simulations. The horizontal solid lines indicate true parameter values.

effects, as we emphasized in Section S.3. In this paper, we do not formally address the question of inference after selection of the number of groups.

S.7. INCOME AND (WAVES OF) DEMOCRACY: ADDITIONAL RESULTS

S.7.1. *Huntington's Theory*

The grouped fixed-effects approach provides a useful complement to fixed-effects in order to study the observed and unobserved determinants of democracy. A conceptual motivation for the GFE model can be found in Samuel Huntington's influential work on the "third wave" of democratization. Huntington (1991) emphasized the importance of international and regional factors as drivers of transitions to democracy and autocracy, resulting in groups of countries making transitions at similar points in time; that is, in "waves" of democratization. Huntington distinguished three waves of democratization: the first one starting in the 1820s in the United States and ending with World

TABLE S.IX

MEAN AND STANDARD DEVIATION OF GFE WHEN G IS ESTIMATED[a]

	θ_1 (Coeff. $y_{i,t-1}$)			θ_2 (Coeff. $\widetilde{x}_{it}$)			$\frac{\theta_2}{1-\theta_1}$		
	True	GFE	Std.	True	GFE	Std.	True	GFE	Std.
$G^0 = 3$	0.407	0.392	0.047	0.089	0.101	0.015	0.151	0.166	0.023
$G^0 = 10$	0.277	0.281	0.048	0.075	0.079	0.011	0.104	0.110	0.014

[a]Model (12) with G^0 groups. The columns labeled "GFE" and "Std." refer to the mean and standard deviation of GFE parameter estimates across 100 simulations, respectively. The number of groups G is selected according to BIC in every simulation. Errors are i.i.d. normal.

War I, the second wave lasting between the end of World War II and the early 1960s, and the third wave starting with the Portuguese revolution in 1974. The first two waves were followed by two "counterwaves," in the 1930s and the 1960s, respectively.

Along with other examples, Huntington mentioned the influence of the U.S. administration in the 1970s and changes in the Soviet Union in the early 1980s, the influence of the European Union in the late 1970s, or changes in the Catholic Church following the second Vatican council, as possible drivers of the clustering of transitions towards democracy that occurred between 1974 and 1990. Huntington's arguments are consistent with the grouped fixed-effects model: for example, $g_i = g$ could denote being predominantly Catholic, and α_{gt} could be the effect of the influence of the Catholic Church on the political evolution of the country. However, our estimation framework is agnostic about the causes of the "waves" of democracy, as it recovers heterogeneous patterns of political evolution from the data.

S.7.2. *Complements to the Main Estimation Results*

Replicating Acemoglu et al. (2008)

Table S.X replicates the main specification from Acemoglu et al. given by equation (22) in the paper, with $\eta_i + \delta_t$ instead of $\alpha_{g_i t}$. Democracy is measured according to the Freedom House indicator, and log-GDP per capita is taken from the Penn World tables. All data are taken at the five-year frequency. We consider two different samples: a balanced panel, which covers 90 countries on the period 1970–2000, and an unbalanced panel, which covers 150 countries on the period 1960–2000. According to the pooled OLS regressions, and regardless of the sample used, there is a statistically significant association between income and democracy. The point estimates of the cumulative income effect $\theta_2/(1-\theta_1)$ imply that a 10% increase in income per capita is associated with a 2.5% increase in the Freedom House score.[28] However, in both data sets, the fixed-effects estimates of the income coefficient are small or negative, and insignificant from zero.

Coefficients of Income and Lagged Democracy

Table S.XI reports three standard error estimates (all of them clustered at the country level): based on a large-T normal approximation, based on Pollard's (1982) fixed-T normal approximation, and based on the bootstrap (our more conservative estimates, shown in Figure 1 in the paper). According to

[28]To assess the magnitude of this effect, note that the Freedom House measure is normalized to lie between zero and 1, and that its mean and standard deviation in the balanced sample are 0.55 and 0.37, respectively.

TABLE S.X

INCOME AND DEMOCRACY, OLS AND FIXED-EFFECTS[a]

	Unbalanced Panel		Balanced Panel	
	(1)	(2)	(3)	(4)
Lag democracy (θ_1)	0.706	0.379	0.665	0.283
	(0.035)	(0.051)	(0.049)	(0.058)
Income (θ_2)	0.072	0.010	0.083	−0.031
	(0.010)	(0.035)	(0.014)	(0.049)
Cumulative income ($\frac{\theta_2}{1-\theta_1}$)	0.246	0.017	0.246)	−0.044
	(0.031)	(0.056)	(0.019)	(0.069)
Observations	945	945	630	630
Countries	150	150	90	90
R-squared	0.725	0.796	0.721	0.799
Time dummies	Yes	Yes	Yes	Yes
Country fixed-effects	No	Yes	No	Yes

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

our estimates, the cumulative income effect is statistically significant.[29] However, it is quantitatively small: only 40% of the pooled OLS estimate when $G \geq 5$. Moreover, we will see in the next subsection that the association between income and democracy disappears in a specification that combines both time-varying grouped effects and time-invariant country-specific effects.

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

Another result of Table S.XI is that $G = 10$ is optimal according to BIC. Recall from Section S.3 that this criterion provides an upper bound on the true number of groups if T grows at a slower rate than N. Note also that the GFE estimates in Figure 1 do not vary much between $G = 5$ and $G = 15$. According to the discussion in Section S.3, this is consistent with the true number of groups being actually *smaller* than 10. Optimal choice of G in practice is a no-

[29]Note that the within-group (i.e., within-$(\widehat{g}_i, t)$) variance of income remains sizable as the number of groups increases: it is 65% of the total income variance when $G = 3$, 48% when $G = 10$, and still 43% when $G = 15$. This is substantially larger than the within-country variance of income (6%). In contrast, the within-group variance of democracy is 10% when $G = 15$, whereas the within-country variance is 26%. This difference arises because the groups are estimated in order to fit the outcome (democracy), but not necessarily the regressor (income).

TABLE S.XI

INCOME AND DEMOCRACY, GFE ESTIMATES[a]

G	Objective	BIC	Lag. Dem. (θ_1)	Income (θ_2)	Cum. Income ($\frac{\theta_2}{1-\theta_1}$)
1	24.301	0.052	0.665	0.083	0.247
			(0.049)	(0.014)	(0.018)
2	19.847	0.046	0.601	0.061	0.152
			(0.041, 0.061, 0.072)	(0.011, 0.013, 0.019)	(0.021, 0.030, 0.058)
3	16.599	0.042	0.407	0.089	0.151
			(0.052, 0.083, 0.129)	(0.011, 0.015, 0.019)	(0.013, 0.022, 0.036)
4	14.319	0.039	0.302	0.082	0.118
			(0.054, 0.108, 0.140)	(0.009, 0.012, 0.017)	(0.011, 0.021, 0.038)
5	12.593	0.037	0.255	0.079	0.107
			(0.050, 0.088, 0.134)	(0.010, 0.012, 0.015)	(0.009, 0.014, 0.040)
6	11.132	0.036	0.465	0.064	0.119
			(0.043, 0.054, 0.122)	(0.007, 0.008, 0.012)	(0.011, 0.014, 0.030)
7	10.059	0.035	0.403	0.065	0.108
			(0.043, 0.074, 0.117)	(0.008, 0.013, 0.013)	(0.011, 0.019, 0.027)
8	9.251	0.035	0.333	0.070	0.104
			(0.044, 0.085, 0.122)	(0.008, 0.012, 0.013)	(0.010, 0.014, 0.033)
9	8.426	0.034	0.312	0.069	0.101
			(0.045, 0.072, 0.123)	(0.008, 0.010, 0.013)	(0.010, 0.011, 0.031)
10*	7.749	0.034	0.277	0.075	0.104
			(0.049, 0.062, 0.124)	(0.008, 0.010, 0.015)	(0.009, 0.011, 0.034)
11	7.218	0.034	0.293	0.073	0.104
			(0.042, 0.062, 0.130)	(0.008, 0.012, 0.014)	(0.009, 0.013, 0.030)
12	6.809	0.034	0.304	0.074	0.107
			(0.044, 0.054, 0.109)	(0.008, 0.009, 0.015)	(0.009, 0.010, 0.037)
13	6.391	0.035	0.236	0.072	0.094
			(0.040, 0.046, 0.120)	(0.009, 0.010, 0.014)	(0.009, 0.010, 0.031)
14	5.996	0.035	0.237	0.071	0.094
			(0.042, 0.047, 0.119)	(0.009, 0.010, 0.017)	(0.009, 0.010, 0.038)
15	5.664	0.035	0.244	0.071	0.094
			(0.043, 0.046, 0.127)	(0.009, 0.010, 0.015)	(0.009, 0.010, 0.040)
Fixed-effects	17.517	–	0.284	−0.031	−0.044
			(0.058)	(0.049)	(0.069)

[a]See the notes to Figure 1 in the paper. The table reports the value of the objective function, the Bayesian information criterion, and GFE coefficient estimates with their standard errors for various values of the number of groups G. Three different standard error estimates (clustered at the country level) are shown in parentheses: based on the large-T normal approximation, on Pollard's (1982) fixed-T normal approximation, and on the bootstrap, respectively. Computation using Algorithm 2 (5; 10; 5). The parameter $\widehat{\sigma}^2$ in BIC was computed using $G_{\max} = 15$. The last row in the table shows the same figures for fixed-effects regression.

toriously difficult problem in related contexts (e.g., mixture and factor models), which deserves further study.

Lastly, the implied cumulative effect of income shown in Figure 1 is almost identical to the estimated income effect when using a specification that only

controls for lagged GDP per capita and does not include lagged democracy (not reported).

Grouped Patterns

In addition to the group-specific means shown in Figure 2 in the paper, Figure S.3 reports uniform 50%-confidence bands for both Freedom House score and lagged log-GDP per capita (thick dashed-dotted lines) for each of the four estimated groups.[30] The figure also shows all country paths of democracy and income over time (thin dotted lines). The left panel shows that, within each group, most countries tend to follow a common group pattern of democracy.[31] At the same time, however, there is evidence of a substantial amount of heterogeneity in democracy paths, which is only imperfectly captured using the parsimonious 4-groups model. In the next subsection, we will present estimates that allow for additional, within-group heterogeneity.

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

Although the estimated groups exhibit a strong spatial clustering, they do not match a simple geographic division. To illustrate this, we report in Figure S.5 the group-specific time effects and averages of democracy and income, respectively, when the continents are used to form five groups. The results show that, although this simple geographic division yields a clear separation in terms of income and democracy levels, the time patterns are not as clearly separated as in Figure 2. In particular, this specification is not able to distinguish between stable and transition patterns within South America or Africa. In contrast, the grouped fixed-effects estimator selects the grouping that maximizes between-group variation, leading to better identification of stable and transition patterns.

As a different strategy, one could use external data to attempt to classify countries. This is the approach taken by Papaioannou and Siourounis (2008),

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

[31]As a complement, Table S.XII reports the 1970–2000 evolution of a binary measure of democracy, which classifies as "democratic" (resp., "nondemocratic") a country whose Freedom House score is strictly higher (resp., lower) than 0.50.

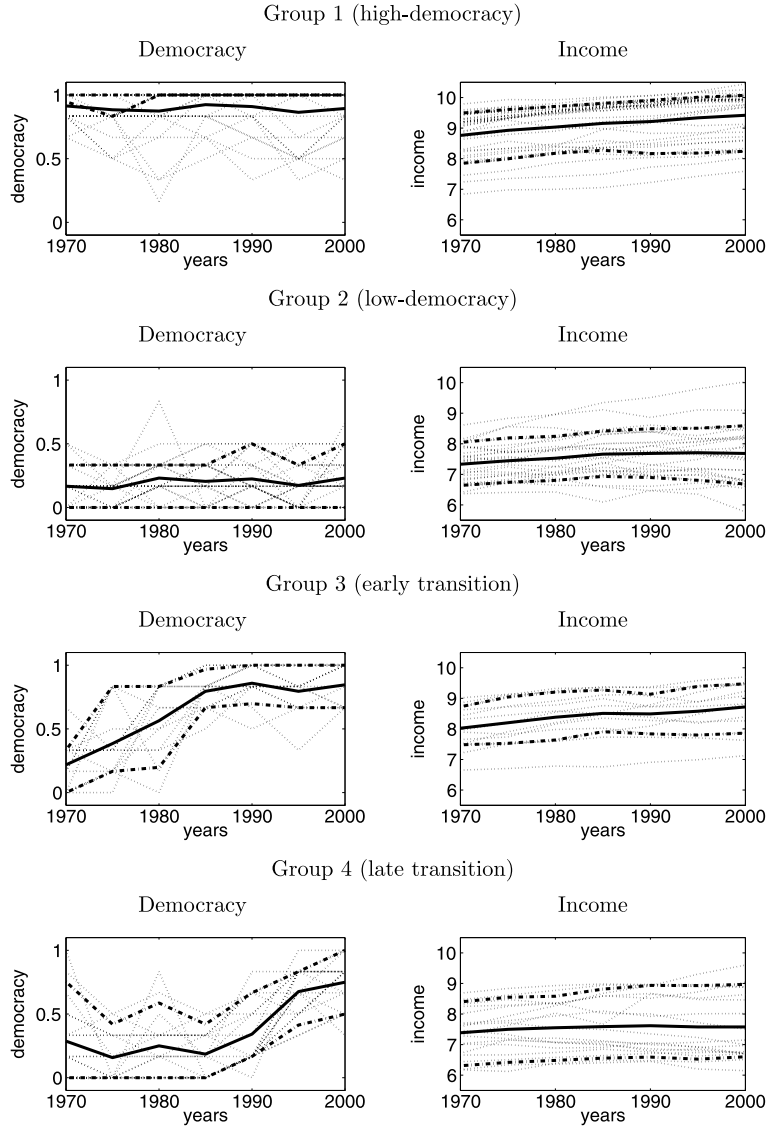

FIGURE S.3.—Confidence bands and data paths of democracy and income ($G = 4$). *Note*: See the notes to Figure 2 in the paper. The left column shows the mean normalized Freedom House score (thick solid lines), a uniform 50%-confidence band (thick dashed-dotted lines), as well as the plot of all democracy paths in the data (thin dotted lines), by group. The right column shows the same figures for lagged log-GDP per capita.

TABLE S.XII

BINARY MEASURE OF DEMOCRACY, TRANSITIONS 1970/2000 BY GROUP ($G = 4$)[a]

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

[a]See the notes to Table S.XI. Here we code as "nondemocratic" (i.e., 0) countries whose Freedom House score is lower than 0.50, and as "democratic" (i.e., 1) countries with a score > 0.50. The numbers a/b in the table denote transition from state $a \in \{0, 1\}$ in 1970 to state $b \in \{0, 1\}$ in 2000.

who combined electoral archives and historical resources for this purpose. Interestingly, their classification of the type of political evolution closely matches the results of GFE estimation. One of the few clear differences between the classification in Papaioannou and Siourounis (2008) and ours is Iran, which is consistently classified as a "low democracy" country according to our results (e.g., in Group 2), while they classify it as a "borderline" democratization case. Note that, unlike this data-intensive approach, our automatic method does not require the use of external data.

S.7.3. *Additional Specifications*

We next summarize the results corresponding to several additional specifications.

Unbalanced Panel

First, we use the unbalanced panel that covers the period 1960–2000. After dropping all countries with fewer than three observations, we obtain an unbalanced sample of 118 countries.[32] See Section S.5.4 for a description of GFE in unbalanced panels. The cumulative income effect is close to the one that we estimated on the balanced sample: for example, it is 0.13 for $G = 4$ and 0.12 for $G = 10$. Interestingly, the group classification is very similar between the two samples: when $G = 4$, the group-specific patterns also highlight high- and low-democracy countries, as well as early and late transition countries. Moreover, out of the 90 countries of the balanced sample, only 6 change groups when estimated on the unbalanced panel. All the countries whose group changes switch from "late" to "early" transition. For example, Mexico, Philippines, and Taiwan become part of the early transition countries. As for those countries that are

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

TABLE S.XIII

GROUP MEMBERSHIP ESTIMATES, VARIOUS SPECIFICATIONS[a]

	Model (1)					Model (S.40)		Model (5)
Country	$G=2$	$G=3$	$G=4$	$G=5$	$G=6$	$\{G_1, G_2\}=\{3,(5,2,2)\}$		$G=3$
Algeria	2	2	2	2	2	Stable	Low	Stable
Argentina	1	3	3	3	3	Early	Low	Early
Australia	1	1	1	1	1	Stable	High	Stable
Austria	1	1	1	1	1	Stable	High	Stable
Belgium	1	1	1	1	1	Stable	High	Stable
Benin	2	3	4	4	4	Late	Low	Late
Bolivia	1	3	3	3	3	Early	Low	Late
Brazil	1	3	3	3	3	Early	Low	Early
Burkina Faso	1	1	4	5	5	Stable	Medium-Low	Stable
Burundi	2	2	2	2	2	Stable	Low	Stable
Cameroon	2	2	2	2	2	Stable	Low	Stable
Canada	1	1	1	1	1	Stable	High	Stable
Central African Rep.	2	3	4	4	4	Late	Low	Late
Chad	2	2	2	2	2	Stable	Low	Stable
Chile	1	3	4	5	5	Late	High	Late
China	2	2	2	2	2	Stable	Low	Stable
Colombia	1	1	1	1	1	Stable	Medium-High	Stable
Congo, Dem. Rep.	2	2	2	2	2	Stable	Low	Stable
Congo Republic	2	2	2	2	2	Stable	Low	Stable
Costa Rica	1	1	1	1	1	Stable	High	Stable
Cote d'Ivoire	2	2	2	2	2	Stable	Low	Stable
Cyprus	1	1	1	1	1	Stable	Medium-High	Late
Denmark	1	1	1	1	1	Stable	High	Stable
Dominican Republic	1	1	1	1	1	Stable	Medium-High	Stable
Ecuador	2	3	3	3	6	Early	Low	Early
Egypt	2	2	2	2	2	Stable	Medium-Low	Stable
El Salvador	1	1	1	1	3	Stable	Medium-High	Stable
Finland	1	1	1	1	1	Stable	Medium-High	Stable
France	1	1	1	1	1	Stable	High	Stable
Gabon	2	2	2	2	2	Stable	Low	Stable
Ghana	2	3	4	4	6	Late	High	Late
Greece	2	3	3	3	3	Early	High	Early
Guatemala	1	1	1	5	5	Stable	Medium	Stable
Guinea	2	2	2	2	2	Stable	Low	Stable
Honduras	2	3	3	3	3	Early	Low	Early
Iceland	1	1	1	1	1	Stable	High	Stable
India	1	1	1	1	1	Stable	High	Stable
Indonesia	1	2	2	5	5	Stable	Medium-Low	Stable
Iran	2	2	2	2	2	Stable	Low	Stable
Ireland	1	1	1	1	1	Stable	High	Stable
Israel	1	1	1	1	1	Stable	Medium-High	Stable
Italy	1	1	1	1	1	Stable	High	Stable

(*Continues*)

TABLE S.XIII—*Continued*

	Model (1)					Model (S.40)		Model (5)
Country	$G=2$	$G=3$	$G=4$	$G=5$	$G=6$	$\{G_1, G_2\}=\{3,(5,2,2)\}$		$G=3$
Jamaica	1	1	1	1	1	Stable	High	Stable
Japan	1	1	1	1	1	Stable	High	Stable
Jordan	2	2	2	2	2	Stable	Medium-Low	Late
Kenya	2	2	2	2	2	Stable	Medium-Low	Stable
Korea, Rep.	1	3	3	3	3	Early	Low	Late
Luxembourg	1	1	1	1	1	Stable	High	Stable
Madagascar	2	3	4	4	4	Late	High	Late
Malawi	2	3	4	4	4	Late	Low	Late
Malaysia	1	1	1	5	1	Stable	Medium	Stable
Mali	2	3	4	4	4	Late	Low	Late
Mauritania	2	2	2	2	2	Stable	Low	Stable
Mexico	2	2	4	5	6	Stable	Medium	Stable
Morocco	1	2	2	5	2	Stable	Medium-Low	Stable
Nepal	1	1	3	3	1	Early	Low	Early
Netherlands	1	1	1	1	1	Stable	High	Stable
New Zealand	1	1	1	1	1	Stable	High	Stable
Nicaragua	1	3	4	5	5	Stable	Medium	Stable
Niger	2	3	4	4	4	Late	Low	Late
Nigeria	2	2	2	5	6	Stable	Medium-Low	Stable
Norway	1	1	1	1	1	Stable	High	Stable
Panama	2	3	4	4	6	Late	Low	Late
Paraguay	1	2	2	5	5	Stable	Medium-Low	Stable
Peru	2	2	3	3	6	Early	Low	Early
Philippines	2	3	4	3	4	Late	High	Late
Portugal	1	1	3	3	1	Early	High	Early
Romania	2	3	4	4	4	Late	Low	Late
Rwanda	2	2	2	2	2	Stable	Low	Stable
Sierra Leone	2	2	2	5	5	Stable	Medium-Low	Stable
Singapore	2	2	2	2	2	Stable	Low	Stable
South Africa	1	3	4	4	4	Late	High	Late
Spain	1	1	3	3	1	Early	High	Early
Sri Lanka	1	1	1	1	1	Stable	Medium-High	Stable
Sweden	1	1	1	1	1	Stable	High	Stable
Switzerland	1	1	1	1	1	Stable	High	Stable
Syria	2	2	2	2	2	Stable	Low	Stable
Taiwan	2	3	4	4	5	Late	High	Late
Tanzania	2	3	4	4	4	Stable	Medium-Low	Stable
Thailand	1	1	3	3	3	Early	High	Early
Togo	2	2	2	2	2	Stable	Low	Stable
Trinidad and Tobago	1	1	1	1	1	Stable	High	Stable
Tunisia	2	2	2	2	2	Stable	Low	Stable
Uganda	2	2	2	2	2	Stable	Medium-Low	Stable
United Kingdom	1	1	1	1	1	Stable	High	Stable
United States	1	1	1	1	1	Stable	High	Stable

(*Continues*)

TABLE S.XIII—*Continued*

	Model (1)					Model (S.40)		Model (5)
Country	$G=2$	$G=3$	$G=4$	$G=5$	$G=6$	$\{G_1, G_2\} = \{3, (5, 2, 2)\}$		$G=3$
Uruguay	1	3	3	3	3	Early	High	Late
Venezuela	1	1	1	1	1	Stable	Medium-High	Stable
Zambia	2	3	4	4	4	Stable	Medium-Low	Stable

[a]Group membership, on the balanced panel from Acemoglu et al. (2008). Columns 2 to 6 show the GFE estimates based on the baseline model, for $G = 2, \ldots, 6$. The next two columns show estimates from a two-layer specification, with $G_1 = 3$ ("Stable," "Early," and "Late," respectively), and $G_2 = \{5, 2, 2\}$ ("High" and "Low," with "Medium-High," "Medium," and "Medium-Low" as intermediate categories for stable countries). The last column shows GFE estimates in deviations to country-specific means, for $G = 3$; see equation (S.21).

not in the balanced sample: Haiti and Zimbabwe are classified in Group 2 (low-democracy), Poland and Hungary in Group 4 (late transition), and Botswana is classified in Group 1 (high-democracy).

Measure of Democracy

As a second exercise, we follow Acemoglu et al. (2008) and use a different measure of democracy: the (normalized) composite Polity index. The balanced panel contains 75 countries, for the same time periods. The grouped fixed-effects estimates are similar to the results obtained using the Freedom House measure. The income effect is 0.20 in the pooled OLS regression, 0.06 for GFE with $G = 2$, and decreases slightly to 0.05 when $G = 15$, significant. Moreover, time patterns and country classification are also similar, although there are some differences related to the measurement of democracy. For example, for $G = 4$, group membership coincides with the one shown in Table S.XIII except in 11 cases. One of the major disagreements between the two sets of results is South Africa, whose 1980 Polity index is 0.70, while its Freedom House score is 0.33.

Additional Covariates

As a third exercise, we include additional controls in model (22). Specifically, following Acemoglu et al. (2008), we control for education, log-population size, and age group percentages (five categories, plus median age). The results are very similar to the main specification. When controlling for education and population size only, the income effect has a similar magnitude (≈ 0.10, significant), while when adding age structure as a control, the cumulative income effect drops to 0.05, marginally significant. For both specifications, the time patterns and country classification documented in Figure 2 in the paper remain almost unchanged.[33]

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

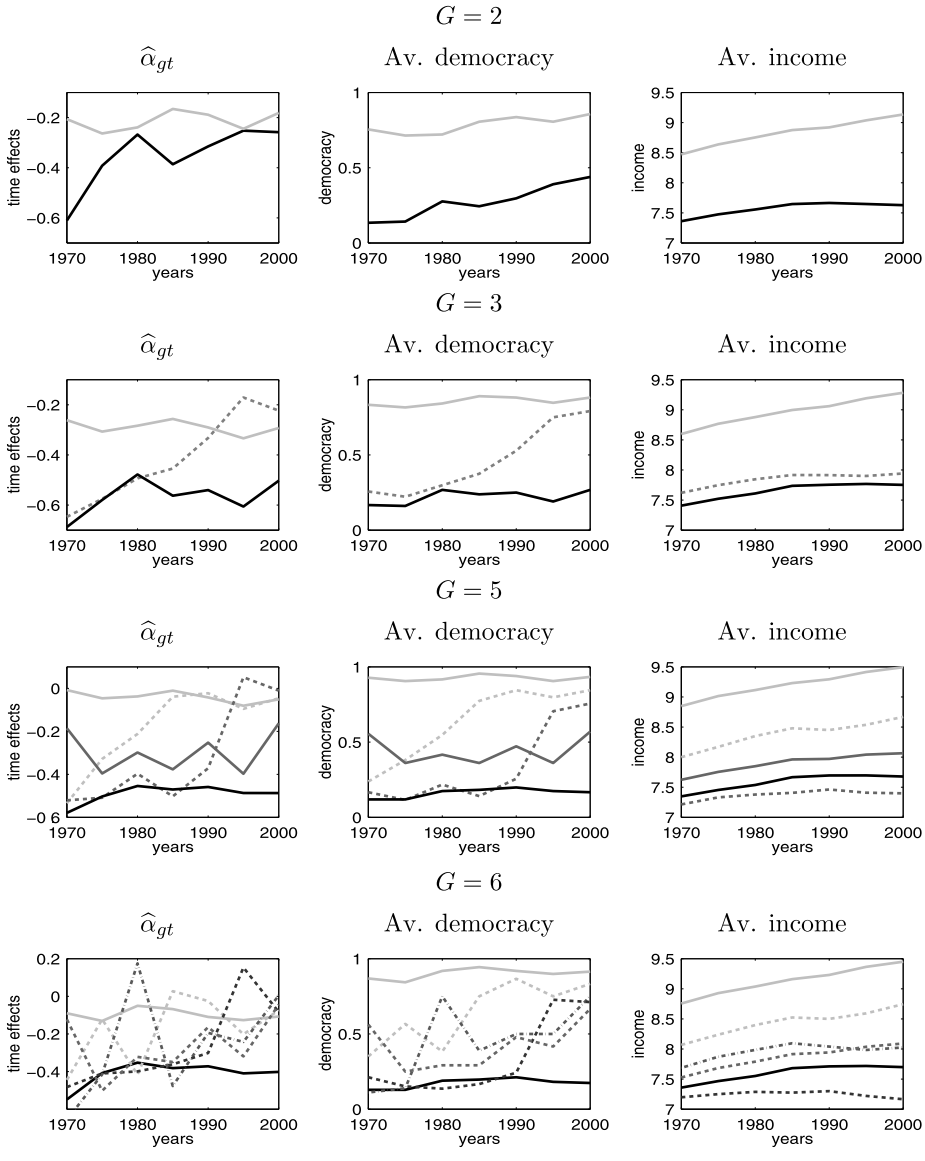

FIGURE S.4.—Patterns of heterogeneity, various G. *Note*: See the notes to Figure 1 in the paper. The left column reports the group-specific time effects $\widehat{\alpha}_{gt}$ for $G = 2$, $G = 3$, $G = 5$, and $G = 6$, from top to bottom. The other two columns show the group-specific averages of democracy and lagged log-GDP per capita, respectively. Calendar years (1970–2000) are shown on the x-axis.

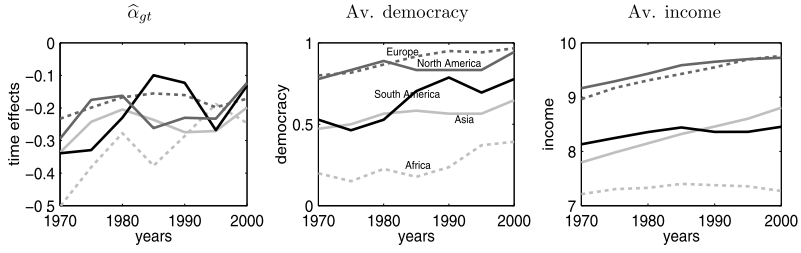

FIGURE S.5.—Continent-specific time-effects. *Note*: See the notes to Figure S.4. The five groups are Europe, North-America (including Mexico), South-America, Asia (including Australia and New-Zealand), and Africa.

Grouped Patterns and Country Fixed-Effects

As a fourth and important exercise, we show the results of a model that combines time-varying group-specific effects and time-invariant country-specific effects, as in equation (5) in the paper. The model is estimated using grouped fixed-effects in deviations to country-specific means; see equation (S.21). Table S.XIV shows the estimates of the coefficients of income and lagged democracy. According to these results, the implied cumulative effect of income on democracy is insignificant, in contrast with the quantitatively small but statistically significant effect obtained using baseline GFE (see Figure 1 in the paper). In Table S.XV, we report Instrumental Variables estimates in first differences, using group membership estimates based on GFE in deviations to country-specific means, and using the second lag of democracy as instrument. These

TABLE S.XIV

INCOME AND DEMOCRACY, GFE ESTIMATES WITH COUNTRY-SPECIFIC FIXED-EFFECTS[a]

G	Objective	Lag. Dem. (θ_1)	Income (θ_2)	Cum. Income ($\frac{\theta_2}{1-\theta_1}$)
1	17.517	0.284	−0.031	−0.044
		(0.058)	(0.049)	(0.069)
2	12.859	0.061	−0.038	−0.040
		(0.049)	(0.027)	(0.029)
3	10.400	−0.033	−0.035	−0.034
		(0.043)	(0.027)	(0.027)
4	9.221	−0.072	0.045	0.042
		(0.046)	(0.027)	(0.025)
5	8.174	−0.093	−0.013	−0.011
		(0.042)	(0.026)	(0.024)

[a]See the notes to Table S.XI. The table reports GFE estimates in deviations to country-specific means (i.e., net of country fixed-effects); see equation (S.21). Clustered standard errors based on the large-T normal approximation in parentheses.

TABLE S.XV

INCOME AND DEMOCRACY, INSTRUMENTAL VARIABLES ESTIMATES[a]

G	Lag. Dem. (θ_1)	Income (θ_2)	Cum. Income ($\frac{\theta_2}{1-\theta_1}$)
1	0.472	−0.075	−0.142
	(0.131)	(0.068)	(0.114)
2	0.338	−0.063	−0.095
	(0.120)	(0.062)	(0.105)
3	0.202	−0.064	−0.080
	(0.094)	(0.055)	(0.084)
4	0.065	−0.037	−0.040
	(0.089)	(0.049)	(0.085)
5	0.085	−0.078	−0.085
	(0.089)	(0.049)	(0.081)

[a]See the notes to Table S.XI. The table reports IV estimates in first differences, using group estimates based on GFE in deviations to country-specific means (i.e., net of country fixed-effects) and using the second lag of democracy as instrument; see equation (S.30). Clustered standard errors based on the large-T normal approximation in parentheses.

estimates are computed using (S.30). The estimates of the coefficient of lagged democracy are larger than in Table S.XIV, consistently with the intuition that IV corrects for downward small-T bias. Moreover, the point estimates of the income effect are negative, and are always insignificant at conventional levels. Both Tables S.XIV and S.XV thus show that, when estimated using GFE estimators that allow for time-invariant fixed-effects and time-varying grouped patterns at the same time, the income effects are in line with the fixed-effects estimate.

However, the estimated time patterns are remarkably robust to the inclusion of country fixed-effects. Under the conditions spelled out in Section S.4, our approach allows to consistently estimate group membership even in the presence of country-specific fixed-effects. The upper panel in Figure S.6 shows that a specification allowing for three different types of time patterns in addition to the country-specific fixed-effects yields a similar division between "stable," "early transition," and "late transition" countries. Moreover, the last column in Table S.XIII shows that the match with the classification without country fixed-effects and $G = 4$ is perfect for 80 out of the 90 countries, the "stable" group mostly comprising countries that belonged to Groups 1 and 2 in the baseline specification (see Figure 2 in the paper). We also estimated the model without including lagged democracy as a control and found very similar results. Indeed, similar time profiles and group classifications emerge when using the standard *kmeans* algorithm (without covariates), in levels or in deviations to country-specific means.

We also estimated the model in first differences (not reported). One issue with the first-differenced data is the presence of a mass point at zero for almost 60% of observations in the balanced panel when using the Freedom House

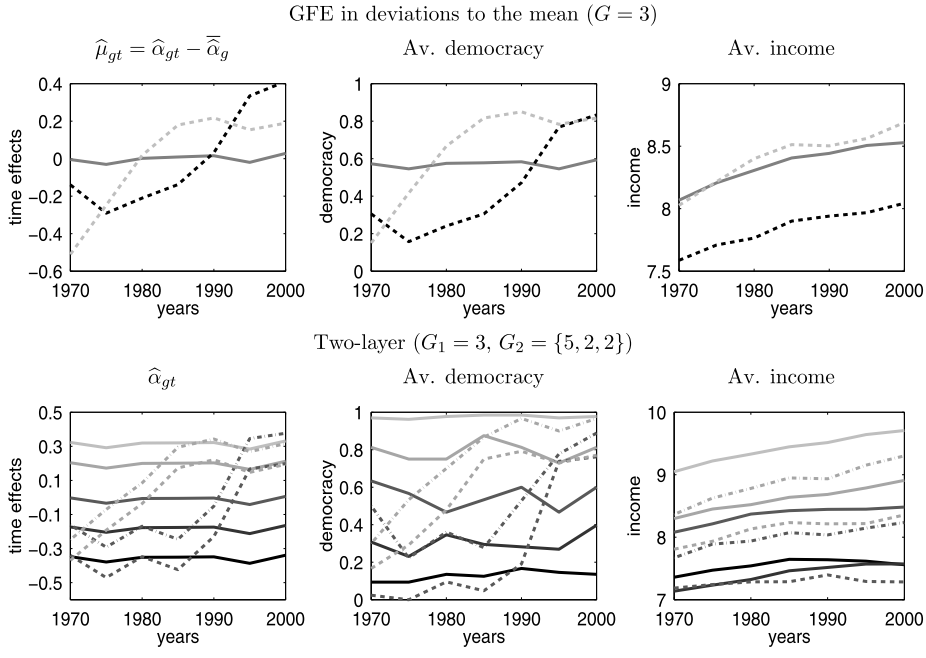

FIGURE S.6.—Grouped patterns and time-invariant heterogeneity. *Note*: See the notes to Figure S.4. The top panel shows the results of GFE estimation in deviation to country-specific means; see equation (S.21). The bottom panel shows the results of the two-layer specification (S.40).

measure of democracy. Although the results show some discrepancies with our baseline group classification, particularly for the early transition group, they similarly highlight the presence of three types of time profiles: stable, early, and late transition.

As a related exercise, we experiment with the two-layer model of unobserved heterogeneity (S.40). This model has G_1 groups with time-varying patterns, and within each of these groups it has G_2 subgroups whose time patterns differ from the common one by an intercept shift. The two-layer model is more parsimonious than model (5), and it may be well-suited given the short length of the panel. We allow for a different number of subgroups within each group, and assume the following two-layer grouped structure:

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

The lower panel of Figure S.6 shows the time-varying group-specific patterns, and the next-to-last two columns in Table S.XIII show group membership by country. We see that the two-layer model delivers a clear separation

between stable countries, early transition countries, and late transition countries. This output is similar to the baseline GFE specification with $G = 4$, and to the estimates in deviations to country-specific means. Note that the two-layer specification and the latter one deliver almost identical group classifications (except in five cases).

In addition, the results provide evidence that the three time-varying groups are heterogeneous themselves. Stable countries show the highest degree of heterogeneity, with five subgroups: high-democracy countries (such as the United States, Japan, Western Europe), medium-high-democracy (Colombia, Venezuela), intermediate (Turkey, Malaysia), medium-low (Paraguay, Indonesia, Egypt), and low-democracy countries (China, Iran). Early transition countries are divided into high (Spain, Portugal) and low (part of Latin America) democracy levels. Similarly, late transition countries are also divided into high (South Africa, Panama) and low (part of Sub-Saharan Africa). Note that the fact that stable countries are separated into five subgroups, whereas early and late transition countries are divided into two subgroups each, is a result of estimation, not of modeling assumptions. At the same time, the division into $(5, 2, 2)$ groups is an assumption. Optimal choice of the number of groups in this context is an interesting question that we do not address here.

Heterogeneous Coefficients

As a last exercise, we estimate two versions of model (7) with heterogeneous coefficients. In the first version, only the income coefficient is group-specific, while in the second version, the coefficients of income and democracy are both group-specific. For computation we use an extension of Algorithm 1, with 1,000,000 randomly generated starting parameter values. In Table S.XVI, we report the results for $G = 4$. The parameter estimates show some evidence of heterogeneity in income effects across groups. For example, in Group 2 (which empirically corresponds to a group of low-income, low-democracy countries), the income effect is lower while still significant. At the same time, as in the baseline case, allowing for country-specific effects in addition yields insignificant income effects in all groups (not reported). Interestingly, Figure S.7 shows that, in both versions of the heterogeneous coefficients model, the groups of countries have very similar income and democracy evolution as in the baseline results (compare with Figure 2 in the paper). In particular, these two specifications highlight again the presence of "stable" groups, and of "early" and "late" transition groups. The results of the models with heterogeneous coefficients thus further illustrate the robustness of this classification.

S.7.4. *Explaining the Estimated Grouped Patterns*

The country classification shown in Figure 2 in the paper seems to be a robust feature of the democracy/income relationship in the last third of the twentieth century. An important question is then why the estimated time profiles differ

TABLE S.XVI

INCOME AND DEMOCRACY, GFE ESTIMATES WITH HETEROGENEOUS COEFFICIENTS ($G = 4$)[a]

	Lag. Dem. (θ_1)	Income (θ_2)	Cum. Income ($\frac{\theta_2}{1-\theta_1}$)
		Heterogeneous θ_2	
Group 1	0.288	0.103	0.145
	(0.054)	(0.019)	(0.024)
Group 2	0.288	0.047	0.066
	(0.054)	(0.014)	(0.019)
Group 3	0.288	0.087	0.122
	(0.054)	(0.018)	(0.024)
Group 4	0.288	0.082	0.116
	(0.054)	(0.013)	(0.016)
		Heterogeneous θ_1 and θ_2	
Group 1	0.644	0.070	0.195
	(0.077)	(0.019)	(0.031)
Group 2	0.319	0.041	0.061
	(0.113)	(0.014)	(0.020)
Group 3	0.016	0.122	0.124
	(0.081)	(0.022)	(0.021)
Group 4	0.248	0.090	0.120
	(0.097)	(0.018)	(0.015)

[a]See the notes to Table S.XI. The table reports GFE estimates with heterogeneous coefficients based on two versions of model (7); see equation (S.32). Extension of Algorithm 1 with 1,000,000 randomly generated starting parameter values. Group 1 is "high democracy," 2 is "low democracy," 3 is "early transition," and 4 is "late transition." Clustered standard errors based on the large-T normal approximation in parentheses.

across countries. We now attempt to identify factors that explain why these four estimated groups of countries are associated with such different levels and evolution of democracy and income during this period.

The first set of factors we consider are long-run, historical determinants. Following Acemoglu et al. (2008), we use a measure of constraints on the executive at independence, the rationale being that more stringent constraints may be beneficial to embark on a pro-growth, pro-democracy development path. We also consider the date of independence and a measure of log-GDP per capita in 1500[34] as potential long-run determinants. In addition, we consider the initial democracy level (in 1965), as well as two factors that have been emphasized by the "modernization" theory (Lipset (1959)): log-GDP per capita (in 1965), and a measure of education (average years of schooling, in 1970). We also include shares of Catholic and Protestant in the population (in 1980).

[34]We construct this measure as the difference between log-GDP per capita in 2000, and the change in log-GDP per capita between 1500 and 2000 used by Acemoglu et al. (2008).

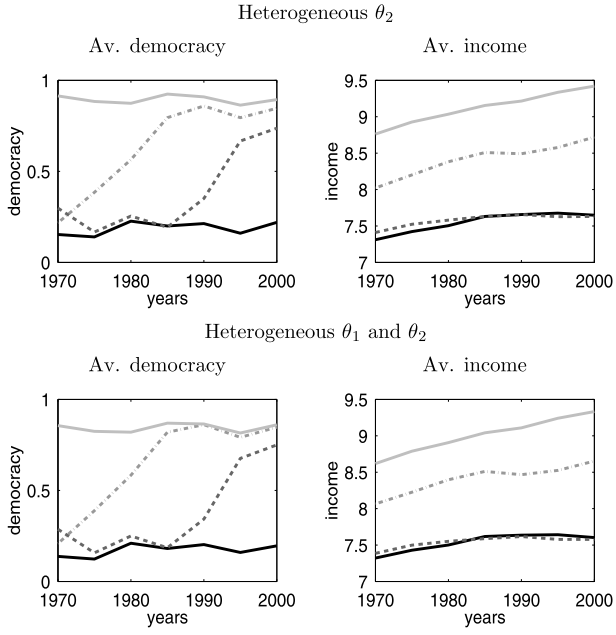

FIGURE S.7.—Grouped patterns, heterogeneous coefficients ($G = 4$). *Note*: See the notes to Figure S.4 and Table S.XVI. GFE estimates based on model (7); see equation (S.32). Averages of democracy and income by group. The top panel shows GFE results in a version of the model where the income coefficient is group-specific. The bottom panel corresponds to a version of the model where the coefficients of income and lagged democracy are both group-specific.

Table S.XVII shows descriptive statistics by group. Both the high-democracy countries (Group 1) and the early transition ones (Group 3) became independent in the nineteenth century on average, while the countries in the two other groups became independent more recently. The high-democracy group had more stringent constraints on the executive at the time of independence. This group also has a higher initial democracy level in 1965,[35] higher initial income and education, and a larger share of Protestant. The early transition group (Group 3) has a higher average education level than the low-democracy group, and a larger share of Catholic (63% versus 23%). Lastly, the late transition group (Group 4) differs little from the low-democracy one in terms of observables, apart from a slightly higher education level.

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

TABLE S.XVII

DESCRIPTIVE STATISTICS, BY GROUP[a]

	Group			
	1 (High Dem.)	2 (Low Dem.)	3 (Early Trans.)	4 (Late Trans.)
log GDP p.c. (1500)	6.52	6.39	6.49	6.30
	(0.300)	(0.437)	(0.141)	(0.236)
Independence Year	1860	1939	1824	1924
	(63.3)	(50.7)	(37.7)	(56.3)
Constraints	0.581	0.258	0.125	0.250
	(0.446)	(0.254)	(0.166)	(0.246)
Democracy (1965)	0.892	0.446	0.510	0.508
	(0.157)	(0.171)	(0.267)	(0.281)
log GDP p.c. (1965)	8.76	7.33	8.02	7.39
	(0.765)	(0.604)	(0.709)	(0.773)
Education (1970)	5.78	1.52	3.63	2.59
	(2.59)	(1.05)	(1.61)	(1.92)
Share Catholic (1980)	0.434	0.232	0.626	0.379
	(0.404)	(0.284)	(0.437)	(0.349)
Share Protestant (1980)	0.248	0.068	0.024	0.140
	(0.330)	(0.088)	(0.032)	(0.160)
Number of observations	33	26	13	18

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

In order to jointly assess the effects of the different factors, we next report in Table S.XVIII the results of multinomial logit regressions of the four estimated groups, using several specifications. In Section S.5.3, we have provided a large-N, T justification for treating the group estimates as data when running the regressions and computing standard errors. The base category is Group 2 (low-democracy). The third row of the top panel of Table S.XVIII shows that constraints on the executive at independence are a significant predictor of the probability of belonging to Group 1 relative to Group 2. This is consistent with the idea that Group 1 and Group 2 countries have embarked on divergent paths at the time of independence, and is suggestive of a very high persistence of early institutions. Note that the effect remains significant at the 10% level even when all other controls (democracy in 1965, income, education…) are included. At the same time, early independence is also associated with a higher likelihood of belonging to Group 1.

However, as shown by the middle and bottom panels of Table S.XVIII, constraints on the executive at independence do not significantly affect the probability of belonging to either of the two transition groups (Groups 3 and 4). This suggests that, while conditions at independence partly explain differences between low- and high-democracy countries, they do not seem to explain the remarkable evolution of transition countries during the recent period.

TABLE S.XVIII

EXPLAINING GROUP MEMBERSHIP[a]

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
					Group 1: High-Democracy (vs. Group 2: Low-Democracy)					
log GDP p.c. (1500)	1.39	0.865	0.698	–	–	–	−0.224	−0.307	−0.465	−0.628
	(0.971)	(1.74)	(1.76)				(2.41)	(2.61)	(2.75)	(2.67)
Independence Year/100	–	−4.55	−4.44	–	–	–	−3.51	−3.72	−3.68	−3.59
		(1.22)	(1.27)				(1.43)	(1.56)	(1.75)	(1.75)
Constraints	–	7.26	7.12	–	–	–	5.67	4.74	4.70	4.52
		(2.00)	(2.06)				(2.49)	(2.60)	(2.62)	(2.77)
Democracy (1965)	–	–	–	7.10	5.80	5.92	–	6.72	6.81	6.24
				(2.11)	(2.56)	(2.66)		(3.39)	(3.44)	(3.65)
log GDP p.c. (1965)	–	–	–	1.51	–	1.09	–	–	0.194	0.447
				(0.587)		(0.883)			(1.25)	(1.35)
Education (1970)	–	–	–	–	0.798	0.492	0.949	0.443	0.418	0.258
					(0.324)	(0.402)	(0.373)	(0.435)	(0.536)	(0.560)
Share Catholic (1980)	–	–	0.611	–	–	–	–	–	–	−0.627
			(1.20)							(1.70)
Share Protestant (1980)	–	–	6.81	–	–	–	–	–	–	3.85
			(4.37)							(6.32)
					Group 3: Early Transition (vs. Group 2: Low-Democracy)					
log GDP p.c. (1500)	0.959	−0.894	−0.504	–	–	–	−1.19	−2.27	−3.48	−3.13
	(1.19)	(1.85)	(1.87)				(2.44)	(2.56)	(3.03)	(2.97)
Independence Year/100	–	−3.53	−3.32	–	–	–	−2.72	−2.96	−4.02	−3.82
		(1.11)	(1.23)				(1.23)	(1.30)	(1.63)	(1.76)
Constraints	–	2.25	2.23	–	–	–	0.939	0.473	0.070	0.010
		(2.10)	(2.34)				(2.47)	(2.56)	(2.57)	(2.95)
Democracy (1965)	–	–	–	−0.232	−1.63	−1.79	–	−1.36	−0.831	−1.37
				(1.69)	(2.03)	(2.08)		(3.03)	(3.02)	(3.16)

(*Continues*)

TABLE S.XVIII—*Continued*

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
log GDP p.c. (1965)	–	–	–	1.40	–	0.503	–	–	−1.87	−1.58
				(0.567)		(0.793)			(1.34)	(1.42)
Education (1970)	–	–	–	–	0.883	0.749	0.570	0.729	1.19	1.18
					(0.311)	(0.357)	(0.361)	(0.425)	(0.565)	(0.601)
Share Catholic (1980)	–	–	1.00	–	–	–	–	–	–	−0.215
			(1.22)							(1.67)
Share Protestant (1980)	–	–	−0.552	–	–	–	–	–	–	−1.55
			(7.87)							(8.93)
				Group 4: Late Transition (vs. Group 2: Low-Democracy)						
log GDP p.c. (1500)	−1.06	−0.968	−0.751	–	–	–	−1.63	−1.97	−1.99	−2.08
	(1.14)	(1.07)	(1.14)				(1.95)	(2.07)	(2.13)	(2.16)
Independence Year/100	–	−0.681	−0.785	–	–	–	−0.027	−0.144	−0.219	−0.007
		(0.635)	(0.763)				(0.926)	(0.939)	(1.03)	(1.38)
Constraints	–	0.485	0.848	–	–	–	−0.607	−1.05	−1.11	−0.527
		(1.30)	(1.39)				(1.74)	(1.86)	(1.88)	(2.22)
Democracy (1965)	–	–	–	1.23	0.047	0.134	–	2.39	2.46	1.50
				(1.43)	(1.93)	(1.89)		(2.45)	(2.45)	(2.77)
log GDP p.c. (1965)	–	–	–	0.021	–	−0.215	–	–	−0.263	0.214
				(0.464)		(0.701)			(0.902)	(1.07)
Education (1970)	–	–	–	–	0.494	0.544	0.597	0.423	0.502	0.331
					(0.302)	(0.349)	(0.358)	(0.389)	(0.439)	(0.476)
Share Catholic (1980)	–	–	0.888	–	–	–	–	–	–	1.20
			(1.19)							(1.90)
Share Protestant (1980)	–	–	5.40	–	–	–	–	–	–	5.23
			(3.87)							(5.78)

[a]Balanced panel from Acemoglu et al. (2008). "Constraints" are constraints on the executive at independence, measured as in Acemoglu, Johnson, and Robinson (2005). Multinomial logit regressions of the estimated groups ($G = 4$). Standard errors clustered at the country level in parentheses. The reference group is Group 2 (low-democracy). Group membership is shown on Figure 2 in the paper. Sample size in the most flexible specification—column (10)—is $N = 68$.

Education positively affects the probability of belonging to Group 3 relative to Group 2, in line with the "modernization" theory. The date of independence also has a positive effect on the likelihood of belonging to the early transition group. These results are consistent with Papaioannou and Siourounis (2008), who modeled the probability of democratization of countries that started the period as autocracies. They found little evidence of an effect of early institutions. In addition, their results also suggest that more educated societies are more likely to become democratic.

In contrast, the bottom panel of Table S.XVIII shows that none of the determinants that we consider (e.g., education or religion) is able to distinguish late transition countries (Group 4) from low-democracy countries (Group 2). Note that most of the late transition countries in Figure 2 are Sub-Saharan African countries, which made democratic transitions in the 1990s. Interestingly, Brückner and Ciccone (2011) documented an association between drought and posterior increases in democracy levels in Sub-Saharan Africa. They interpreted this evidence as suggesting that a fall in *transitory* income may foster democratic change.

Overall, these results point to the need to further study the short- and long-run determinants of political development. Constraints at independence were significantly more stringent in countries that remained democratic between 1970 and 2000, compared to those that remained nondemocratic during the period. However, this measure does not explain why some countries that were nondemocratic at the beginning of the sample period experienced a democratic transition, while others did not. For a sizable share of the world, history appears to have evolved at a fast pace.

APPENDIX

S.A.1. *Proof of Corollary 1*

We have

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

which tends to $\mathcal{N}(0,\Sigma_\theta^{-1}\Omega_\theta\Sigma_\theta^{-1})$ by Assumptions 3(a)–3(c). Result (19) then follows from the fact that $\sqrt{NT}(\widehat{\theta}-\widetilde{\theta})=o_p(1)$.

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

$$\widetilde{\alpha}_{gt} = \frac{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}(y_{it} - x_{it}'\widetilde{\theta})}{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}} \tag{S.46}$$

$$= \alpha_{gt}^0 + \left(\frac{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} x_{it}}{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}} \right)' (\theta^0 - \widetilde{\theta}) + \frac{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} v_{it}}{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}}.$$

Now, using Assumptions 1(b) and 2(a) as well as the above, we have

$$\left(\frac{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} x_{it}}{\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}} \right)' (\theta^0 - \widetilde{\theta}) = O_p\left(\frac{1}{\sqrt{NT}} \right). \tag{S.47}$$

Hence:

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

and (20) follows from a similar argument as before.

This ends the proof of Corollary 1.

S.A.2. *Proof of the Convergence Rate in Equation (21) in the Paper*

To show (21), we will bound the following three quantities in turn:

$$A \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{\alpha}_{\widehat{g}_i t} - \widehat{\alpha}_{g_i^0 t})^2; \quad B \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widehat{\alpha}_{g_i^0 t} - \widetilde{\alpha}_{g_i^0 t})^2;$$

$$C \equiv \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\widetilde{\alpha}_{g_i^0 t} - \alpha_{g_i^0 t}^0)^2.$$

By Assumption 1(a) and Lemma B.4,

$$A = O_p(1)\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\widehat{g}_i \neq g_i^0\} = o_p(T^{-\delta}) \quad \text{for all } \delta > 0.$$

Moreover, by (B.19) in the paper,

$$B \leq \max_{g\in\{1,\ldots,G\}}\frac{1}{T}\sum_{t=1}^{T}(\widehat{\alpha}_{gt} - \widetilde{\alpha}_{gt})^2 = o_p(T^{-\delta}).$$

Lastly, by (S.46), (S.47), and Assumption 2(a),

$$\begin{aligned} C &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{g=1}^{G}\mathbf{1}\{g_i^0 = g\} \\ &\quad \times \left[\left(\frac{\sum_{j=1}^{N}\mathbf{1}\{g_j^0 = g\}x_{jt}}{\sum_{j=1}^{N}\mathbf{1}\{g_j^0 = g\}}\right)'(\theta^0 - \widetilde{\theta}) + \frac{\sum_{j=1}^{N}\mathbf{1}\{g_j^0 = g\}v_{jt}}{\sum_{j=1}^{N}\mathbf{1}\{g_j^0 = g\}}\right]^2 \\ &= O_p\left(\frac{1}{NT}\right) + O_p(1) \times \max_{g\in\{1,\ldots,G\}}\left[\frac{1}{N^2T}\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\mathbf{1}\{g_i^0 = g\}v_{it}\right)^2\right]. \end{aligned}$$

Now, by assumption (see footnote 14 in the paper), we have, for all $g \in \{1,\ldots,G\}$,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{NT}\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\mathbf{1}\{g_i^0 = g\}v_{it}\right)^2\right] \\ &= \frac{1}{NT}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=1}^{T}\mathbb{E}\big(\mathbf{1}\{g_i^0 = g\}\mathbf{1}\{g_j^0 = g\}v_{it}v_{jt}\big) = O(1). \end{aligned}$$

It thus follows that $C = O_p(1/N)$.

Hence, if $N/T^{\nu} \to 0$ for some $\nu > 0$,

$$A + B + C = O_p\left(\frac{1}{N}\right).$$

This proves (21).

S.A.3. *Proof of Proposition S.1*

We start with a lemma.

LEMMA S.1: *We have*

$$\frac{\partial}{\partial\theta'}\mathbb{E}\big[\mathbf{1}\big\{\widehat{g}_i(\theta,\alpha)=g\big\}|x_i=x\big]=\sum_{h\neq g}\left(\int_{S_{gh}}f(y|x)\,dy\right)\frac{(\alpha_h-\alpha_g)'}{\|\alpha_h-\alpha_g\|}x, \tag{S.48}$$

$$\frac{\partial}{\partial\alpha_g'}\mathbb{E}\big[\mathbf{1}\big\{\widehat{g}_i(\theta,\alpha)=g\big\}|x_i=x\big]=\sum_{h\neq g}\left(\int_{S_{gh}}\frac{(y-x\theta-\alpha_g)'}{\|\alpha_h-\alpha_g\|}f(y|x)\,dy\right), \tag{S.49}$$

$$\frac{\partial}{\partial\alpha_{\widetilde{g}}'}\mathbb{E}\big[\mathbf{1}\big\{\widehat{g}_i(\theta,\alpha)=g\big\}|x_i=x\big]=-\left(\int_{S_{g\widetilde{g}}}\frac{(y-x\theta-\alpha_{\widetilde{g}})'}{\|\alpha_{\widetilde{g}}-\alpha_g\|}f(y|x)\,dy\right) \tag{S.50}$$

$$\text{for all } \widetilde{g}\neq g,$$

where S_{gh} *is given by* (S.8).

PROOF: Let

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

Note that

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

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

We have

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

Hence, using differential calculus as in Pollard (1982), we have, for all $k\in\{1,\dots,K\}$,[36]

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

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

[36]Note that, with some abuse of notation, here the integral is relative to the $(T-1)$-dimensional Lebesgue measure.

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

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

where S_{gh} is given by (S.8).

As a result, we have the following identity:

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

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

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

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

Finally, let

$$\begin{aligned} S^*_{gh}(\xi) = \big\{ & y\in\mathbb{R}^T,\ \big\|y - x\theta^*(\xi) - \alpha_g\big\|^2 = \big\|y - x\theta^*(\xi) - \alpha_h\big\|^2, \text{ and} \\ & \big\|y - x\theta^*(\xi) - \alpha_g\big\|^2 \le \big\|y - x\theta^*(\xi) - \alpha_{\widetilde{h}}\big\|^2 \\ & \text{for all } \widetilde{h}\neq(g,h)\big\}. \end{aligned}$$

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

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

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

In the present case, we have

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

Moreover, $y^*(\xi)$ satisfies

$$y^*(\xi) = y + \lambda(\xi)(\alpha_h-\alpha_g), \tag{S.51}$$

where $\lambda(\xi)$ is such that

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

That is,

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

from which we get, as $y \in S_{gh}$,

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

It thus follows that

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

Combining, we get

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

and hence

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

This shows (S.48).

To show (S.49) and (S.50), we proceed similarly. The only difference is the characterization of the velocity. We start by computing $\nu_g(y;\alpha_{gt})$ for $y \in S_{gh}$. To do this, we define $\lambda(\xi)$ as in (S.51), but now $y^*(\xi)$ solves

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

where

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

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

That is,

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

so

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

It thus follows that

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

Combining the results yields (S.49).

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

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

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

where

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

That is,

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

so

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

Following the above steps yields (S.50). *Q.E.D.*

We then have the following result.

LEMMA S.2:

$$\begin{aligned}(\text{S.52}) \quad &\frac{\partial}{\partial \alpha_g'}\bigg|_{(\overline{\theta}, \overline{\alpha})} \mathbb{E}\left[\mathbf{1}\left\{\widehat{g}_i(\theta, \alpha) = g\right\}(y_i - x_i\overline{\theta} - \overline{\alpha}_g) | x_i = x\right] \\ &= \sum_{h \neq g} \left(\int_{\overline{S}_{gh}} \frac{(y - x\overline{\theta} - \overline{\alpha}_g)(y - x\overline{\theta} - \overline{\alpha}_g)'}{\|\overline{\alpha}_h - \overline{\alpha}_g\|} f(y|x)\, dy \right),\end{aligned}$$

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

$$\frac{\partial}{\partial \alpha_{\widetilde{g}}'}\bigg|_{(\overline{\theta},\overline{\alpha})} \mathbb{E}\big[\mathbf{1}\big\{\widehat{g}_i(\theta,\alpha)=g\big\}(y_i - x_i\overline{\theta} - \overline{\alpha}_g)|x_i = x\big] = -\bigg(\int_{\overline{S}_{g\widetilde{g}}} \frac{(y - x\overline{\theta} - \overline{\alpha}_g)(y - x\overline{\theta} - \overline{\alpha}_{\widetilde{g}})'}{\|\overline{\alpha}_{\widetilde{g}} - \overline{\alpha}_g\|} f(y|x)\,dy\bigg). \tag{S.53}$$

The lemma is a simple consequence of Lemma S.1, so its proof is omitted. Lastly, we prove Proposition S.1. We have

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

where we have used (S.48). We also note that, with probability 1,

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

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

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

Hence

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

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

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

This shows (S.9).

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

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

where we have used (S.49) and (S.50).

We then have

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

where we have used (S.52).

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

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

where we have used (S.53).

This ends the proof of Proposition S.1.

S.A.4. *Proof of Proposition S.2*

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

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

$$\begin{aligned}
\text{(S.54)}\quad & \mathbb{E}\bigg[\sum_{t=1}^{T} x_{it}\big(v_{it} + x_{it}'\big(\theta^0 - \overline{\theta}\big)\big) \\
&\qquad + \sum_{t=1}^{T} x_{it}\mathbf{1}\bigg\{\overline{v}_i \leq \overline{x}_i'\big(\overline{\theta} - \theta^0\big) + \frac{\overline{\alpha}_1 + \overline{\alpha}_2}{2} - \alpha^0\bigg\}\big(\alpha^0 - \overline{\alpha}_1\big) \\
&\qquad + \sum_{t=1}^{T} x_{it}\mathbf{1}\bigg\{\overline{v}_i > \overline{x}_i'\big(\overline{\theta} - \theta^0\big) + \frac{\overline{\alpha}_1 + \overline{\alpha}_2}{2} - \alpha^0\bigg\}\big(\alpha^0 - \overline{\alpha}_2\big)\bigg] \\
&= 0, \\
\text{(S.55)}\quad & \mathbb{E}\bigg[\mathbf{1}\bigg\{\overline{v}_i \leq \overline{x}_i'\big(\overline{\theta} - \theta^0\big) + \frac{\overline{\alpha}_1 + \overline{\alpha}_2}{2} - \alpha^0\bigg\}\big(\overline{v}_i + \overline{x}_i'\big(\theta^0 - \overline{\theta}\big) + \alpha^0 - \overline{\alpha}_1\big)\bigg] \\
&= 0, \\
\text{(S.56)}\quad & \mathbb{E}\bigg[\mathbf{1}\bigg\{\overline{v}_i > \overline{x}_i'\big(\overline{\theta} - \theta^0\big) + \frac{\overline{\alpha}_1 + \overline{\alpha}_2}{2} - \alpha^0\bigg\}\big(\overline{v}_i + \overline{x}_i'\big(\theta^0 - \overline{\theta}\big) + \alpha^0 - \overline{\alpha}_2\big)\bigg] \\
&= 0.
\end{aligned}$$

Now, let a_1 and a_2 be the solutions of

$$T\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i \leq \frac{a_1+a_2}{2} - \alpha^0\right\}\left(\overline{v}_i + \alpha^0 - a_1\right)\right] = 0, \tag{S.57}$$

$$T\mathbb{E}\left[\mathbf{1}\left\{\overline{v}_i > \frac{a_1+a_2}{2} - \alpha^0\right\}\left(\overline{v}_i + \alpha^0 - a_2\right)\right] = 0. \tag{S.58}$$

Note that (θ^0, a_1, a_2) satisfies the moment restrictions (S.54)–(S.56) because, as v_{it} and x_{it} are independent of each other, we have

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

where we have used that the sum of the left-hand sides in (S.57) and (S.58) is zero.

Provided the solution to the population moment restrictions (S.54)–(S.56) be unique,[37] it thus follows that

$$(\overline{\theta}, \overline{\alpha}_1, \overline{\alpha}_2) = \left(\theta^0, a_1, a_2\right). \tag{S.59}$$

Hence $\widehat{\theta} \xrightarrow{p} \theta^0$. In addition, it follows from (S.57)–(S.58) and (S.59) that

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

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

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

$$\overline{\alpha}_1 = \alpha^0 + \mathbb{E}(\overline{v}_i | \overline{v}_i \leq 0), \quad \overline{\alpha}_2 = \alpha^0 + \mathbb{E}(\overline{v}_i | \overline{v}_i > 0).$$

[37]Uniqueness of the population minimum is a key ingredient for showing that $(\widehat{\theta}, \widehat{\alpha}) \xrightarrow{p} (\overline{\theta}, \overline{\alpha})$ as N tends to infinity (Pollard (1981)); see Appendix S.A.3. Uniqueness is implicitly assumed in the statement of Proposition S.2.

The final result comes from the normality assumption, as

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

This ends the proof of Proposition S.2.

S.A.5. *Proof of Proposition S.3*

From (S.21) and Assumptions 1(a)–1(c) and 1(d)–1(g) applied to $x_{it} - \overline{x}_i$ and $v_{it} - \overline{v}_i$, and denoting $\mu^0_{gt} = \alpha^0_{gt} - \overline{\alpha}^0_g$, Theorem 1 yields

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

and

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

In the rest of the proof, we closely follow the proof of Theorem 2. It is immediate to show that $d_H(\widehat{\mu}^{FE}, \mu^0) \xrightarrow{p} 0$. We then follow the proof of Lemma B4 to show an analogous result, by replacing x_{it}, α_{gt}, α^0_{gt}, and v_{it} by $x_{it} - \overline{x}_i$, $\alpha_{gt} - \overline{\alpha}_g$, $\alpha^0_{gt} - \overline{\alpha}^0_g$, and $v_{it} - \overline{v}_i$, respectively. Equation (B.6) in the paper thus becomes

$$\begin{aligned}
\text{(S.60)}\qquad \Pr(\widetilde{Z}_{ig} = 1) \leq \sum_{\widetilde{g}\neq g}\Bigg[& \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it} - \overline{x}_i\| \geq \widetilde{M}\right) \\
& + \Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mu^0_{\widetilde{g}t} - \mu^0_{gt}\right)^2 \leq \frac{c^{FE}_{g,\widetilde{g}}}{2}\right) \\
& + \Pr\left(\frac{1}{T}\sum_{t=1}^{T}(v_{it} - \overline{v}_i)^2 \geq \widetilde{M}\right) \\
& + \Pr\Bigg(\sum_{t=1}^{T}\left(\mu^0_{\widetilde{g}t} - \mu^0_{gt}\right)(v_{it} - \overline{v}_i) \leq -T\frac{c^{FE}_{g,\widetilde{g}}}{4} + TC_1\sqrt{\eta}\sqrt{\widetilde{M}} \\
& + TC_2\sqrt{\eta}\widetilde{M} + TC_3\sqrt{\eta}\Bigg)\Bigg],
\end{aligned}$$

where $c^{FE}_{g,\widetilde{g}} = \operatorname{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}(\mu^0_{gt} - \mu^0_{\widetilde{g}t})^2$.

We bound the last three terms on the right-hand side of (S.60), similarly as in the proof of Theorem 2. Start with the second term. By assumption, we have $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[(\mu_{gt}^0-\mu_{\widetilde{g}t}^0)^2]=c_{g,\widetilde{g}}^{FE}$. So for T large enough, we have

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

Hence

$$\begin{aligned}&\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\big(\mu_{\widetilde{g}t}^0-\mu_{gt}^0\big)^2\leq\frac{c_{g,\widetilde{g}}^{FE}}{2}\right)\\&\quad\leq\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\big[\big(\mu_{\widetilde{g}t}^0-\mu_{gt}^0\big)^2-\mathbb{E}\big(\big(\mu_{\widetilde{g}t}^0-\mu_{gt}^0\big)^2\big)\big]\leq-\frac{c_{g,\widetilde{g}}^{FE}}{6}\right).\end{aligned}$$

To simplify the notation, let us denote $a_t=\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0$. We have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\big[\big(\mu_{\widetilde{g}t}^0-\mu_{gt}^0\big)^2-\mathbb{E}\big(\big(\mu_{\widetilde{g}t}^0-\mu_{gt}^0\big)^2\big)\big]\\&\quad=\frac{1}{T}\sum_{t=1}^{T}\big[(a_t-\overline{a})^2-\mathbb{E}\big((a_t-\overline{a})^2\big)\big]\\&\quad=\left[\frac{1}{T}\sum_{t=1}^{T}\big[a_t^2-\mathbb{E}\big(a_t^2\big)\big]\right]-\big[\overline{a}^2-\mathbb{E}\big(\overline{a}^2\big)\big].\end{aligned}\tag{S.61}$$

The first term on the right-hand side of (S.61) can be bounded using Lemma B.5. To bound the second term, note that

$$\begin{aligned}\big|\overline{a}^2-\mathbb{E}\big(\overline{a}^2\big)\big|&=\big|\overline{a}^2-\big[\mathbb{E}(\overline{a})\big]^2-\operatorname{Var}(\overline{a})\big|\\&=\big|\big(\overline{a}+\mathbb{E}(\overline{a})\big)\big(\overline{a}-\mathbb{E}(\overline{a})\big)-\operatorname{Var}(\overline{a})\big|\\&\leq\big|\big(\overline{a}+\mathbb{E}(\overline{a})\big)\big|\times\big|\big(\overline{a}-\mathbb{E}(\overline{a})\big)\big|+\big|\operatorname{Var}(\overline{a})\big|.\end{aligned}$$

Now, a_t is uniformly bounded by Assumption 1(a). Moreover, by Assumptions 2(c)–2(d), $\lim_{T\to\infty}\operatorname{Var}(\overline{a})=0$. Using Lemma B.5 with $z_t=a_t-\mathbb{E}(a_t)$ thus yields that, for any $z>0$, $\Pr(|\overline{a}^2-\mathbb{E}(\overline{a}^2)|\geq z)=o(T^{-\delta})$ for all $\delta>0$. As a result, the second term on the right-hand side of (S.60) is $o(T^{-\delta})$.

The third and fourth terms on the right-hand side of (S.60) are easy to bound. Indeed: $\frac{1}{T}\sum_{t=1}^{T}(v_{it}-\overline{v}_i)^2 \leq \frac{1}{T}\sum_{t=1}^{T} v_{it}^2$. Moreover, denoting as c^{FE} the minimum of $c_{g,\widetilde{g}}^{FE}$ over all $g \neq \widetilde{g}$ and taking

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

the fourth term on the right-hand side of (S.60) is bounded by

$$\begin{aligned}&\Pr\left(\sum_{t=1}^{T}\left(\mu_{\widetilde{g}t}^0-\mu_{gt}^0\right)(v_{it}-\overline{v}_i) \leq -T\frac{c_{g,\widetilde{g}}^{FE}}{8}\right)\\ &\quad=\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0\right)v_{it}-\left(\overline{\alpha}_{\widetilde{g}}^0-\overline{\alpha}_g^0\right)\overline{v}_i \leq -\frac{c_{g,\widetilde{g}}^{FE}}{8}\right).\end{aligned}$$

Lemma B.5, applied to $z_t=(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0)v_{it}$, allows to bound $\frac{1}{T}\sum_{t=1}^{T}(\alpha_{\widetilde{g}t}^0-\alpha_{gt}^0)v_{it}$. Moreover, $|\overline{\alpha}_{\widetilde{g}}^0-\overline{\alpha}_g^0|$ is uniformly bounded, so Lemma B.5 applied to $z_t=v_{it}$ allows to bound $(\overline{\alpha}_{\widetilde{g}}^0-\overline{\alpha}_g^0)\overline{v}_i$. This shows that the fourth term on the right-hand side of (S.60) is $o(T^{-\delta})$. Lastly, the first term on the right-hand side of (S.60) is bounded analogously as in the proof of Theorem 2, using Assumption S.1(b).

The end of the proof is as in the proof of Theorem 2.

Sufficient Conditions for Assumption S.1(b)

Note that, if x_{it} satisfies Assumption 2(e), then it also satisfies Assumption S.1(b), as

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|x_{it}-\overline{x}_i\| \geq M^*\right) \leq \Pr\left(\frac{2}{T}\sum_{t=1}^{T}\|x_{it}\| \geq M^*\right).$$

Moreover, in models where x_{it} contains a lagged outcome, we have the following result.

PROPOSITION S.5: *Consider model* (5). *Suppose that Assumptions* 1(a), 1(c), *and* 2(c)–2(d) *are satisfied. In addition, suppose that* $x_{it}=(y_{i,t-1},\widetilde{x}_{it}')'$, *and* $\theta=(\rho,\theta_1')'$, *where* $|\rho^0|<1$, $\widetilde{x}_{it}$ *satisfy Assumption* 2(e), *and, for all constants* $F_1>0$, $F_2>0$,

$$\sup_{i\in\{1,\ldots,N\}}\Pr\left(|y_{i0}| \geq F_1T\right)=o\left(T^{-\delta}\right) \quad \textit{for all } \delta>0, \tag{S.62}$$

$$\sup_{i\in\{1,\ldots,N\}}\Pr\left(|\eta_i| \geq F_2T\right)=o\left(T^{-\delta}\right) \quad \textit{for all } \delta>0. \tag{S.63}$$

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

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

PROOF: As Assumption 2(e) implies Assumption S.1(b), we have $\sup_{i\in\{1,\dots,N\}} \Pr(\frac{1}{T}\sum_{t=1}^{T}\|\widetilde{x}_{it}-\overline{\widetilde{x}}_i\| \geq M^*) = o(T^{-\delta})$. Moreover,

$$y_{it} = \widetilde{y}_{it} + \frac{1-\big(\rho^0\big)^t}{1-\rho^0}\eta_i,$$

where

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

So, for all i,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}|y_{i,t-1}-\overline{y}_{i,-1}| \\
&\quad\leq \frac{1}{T}\sum_{t=1}^{T}|\widetilde{y}_{i,t-1}-\overline{\widetilde{y}}_{i,-1}| + \frac{1}{T}\sum_{t=1}^{T}\left|\frac{1}{T}\sum_{s=1}^{T}\frac{\big(\rho^0\big)^{s-1}}{1-\rho^0} - \frac{\big(\rho^0\big)^{t-1}}{1-\rho^0}\right||\eta_i| \\
&\quad\leq \frac{1}{T}\sum_{t=1}^{T}|\widetilde{y}_{i,t-1}-\overline{\widetilde{y}}_{i,-1}| + \frac{1}{T}\frac{2}{\big(1-\big|\rho^0\big|\big)^2}|\eta_i|.
\end{aligned}$$

Now, as in the proof of Proposition B.1 in the paper (and using (S.62)), there exists a positive constant M^* such that

$$\begin{aligned}
&\sup_{i\in\{1,\dots,N\}} \Pr\left(\frac{1}{T}\sum_{t=1}^{T}|\widetilde{y}_{i,t-1}-\overline{\widetilde{y}}_{i,-1}| \geq \frac{M^*}{2}\right) \\
&\quad\leq \sup_{i\in\{1,\dots,N\}} \Pr\left(\frac{2}{T}\sum_{t=1}^{T}|\widetilde{y}_{i,t-1}| \geq \frac{M^*}{2}\right) = o\big(T^{-\delta}\big).
\end{aligned}$$

Moreover, for this M^*, (S.63) implies that

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

This concludes the proof of Proposition S.5. *Q.E.D.*

S.A.6. *Proof of Proposition S.4*

Let

$$\widehat{\mathcal{Q}}(\theta,\alpha,\gamma)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\bigl(y_{it}-x_{it}'\theta_{g_i}-\alpha_{g_it}\bigr)^2,$$

and

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

As in the proof of Theorem 1, we start by showing the following uniform convergence result.

LEMMA S.3: *Let Assumptions* 1(a)–1(f) *hold. Then*

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

PROOF: We have

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

The second and third terms on the right-hand side are bounded as in the proof of Lemma A.1. To bound the fourth term, note that

$$\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}v_{it}x_{it}'\theta_{g_i}\right)^2\le\frac{1}{N}\sum_{i=1}^{N}\|\theta_{g_i}\|^2\left\|\frac{1}{T}\sum_{t=1}^{T}v_{it}x_{it}\right\|^2,$$

which is uniformly $o_p(1)$ by Assumptions 1(a) and 1(d). The first term is thus $o_p(1)$, too. *Q.E.D.*

With some abuse of notation, we use $d_H(\theta_1, \theta_2)$ and $d_H(\alpha_1, \alpha_2)$ to denote the Hausdorff distances on $\mathbb{R}^{GK}$ and $\mathbb{R}^{GT}$, respectively, where $K = \dim x_{it}$.[38] We have the following consistency result.

LEMMA S.4: *Suppose that the conditions of Proposition S.4 are satisfied. Then, as N, T tend to infinity,*

$$d_H\big(\widehat{\theta}^{HC}, \theta^0\big) \xrightarrow{p} 0, \quad \text{and} \quad d_H\big(\widehat{\alpha}^{HC}, \alpha^0\big) \xrightarrow{p} 0.$$

PROOF: Let $(\theta, \alpha, \gamma) \in \Theta^G \times \mathcal{A}^{GT} \times \Gamma_G$. Let also $\alpha_g = (\alpha_{g1}, \ldots, \alpha_{gT})'$. We have

$$\begin{aligned}
(\text{S.64}) \quad & \widetilde{\mathcal{Q}}(\theta, \alpha, \gamma) - \widetilde{\mathcal{Q}}\big(\theta^0, \alpha^0, \gamma^0\big) \\
&= \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \big(x_{it}' \big(\theta^0_{g_i^0} - \theta_{g_i}\big) + \alpha^0_{g_i^0 t} - \alpha_{g_i t}\big)^2 \\
&= \sum_{g=1}^{G} \sum_{\widetilde{g}=1}^{G} \begin{pmatrix} \theta^0_g - \theta_{\widetilde{g}} \\ \dfrac{1}{\sqrt{T}} \big(\alpha^0_g - \alpha_{\widetilde{g}}\big) \end{pmatrix}' M(\gamma, g, \widetilde{g}) \begin{pmatrix} \theta^0_g - \theta_{\widetilde{g}} \\ \dfrac{1}{\sqrt{T}} \big(\alpha^0_g - \alpha_{\widetilde{g}}\big) \end{pmatrix} \\
&\geq \sum_{g=1}^{G} \sum_{\widetilde{g}=1}^{G} \widehat{\rho}(\gamma, g, \widetilde{g}) \left[\big\| \theta^0_g - \theta_{\widetilde{g}} \big\|^2 + \frac{1}{T} \sum_{t=1}^{T} \big(\alpha^0_{gt} - \alpha_{\widetilde{g}t}\big)^2 \right] \\
&\geq \sum_{g=1}^{G} \left(\sum_{\widetilde{g}=1}^{G} \widehat{\rho}(\gamma, g, \widetilde{g}) \right) \min_{\widetilde{g} \in \{1, \ldots, G\}} \left[\big\| \theta^0_g - \theta_{\widetilde{g}} \big\|^2 + \frac{1}{T} \sum_{t=1}^{T} \big(\alpha^0_{gt} - \alpha_{\widetilde{g}t}\big)^2 \right] \\
&\geq \sum_{g=1}^{G} \Big(\max_{\widetilde{g} \in \{1, \ldots, G\}} \widehat{\rho}(\gamma, g, \widetilde{g}) \Big) \min_{\widetilde{g} \in \{1, \ldots, G\}} \left[\big\| \theta^0_g - \theta_{\widetilde{g}} \big\|^2 + \frac{1}{T} \sum_{t=1}^{T} \big(\alpha^0_{gt} - \alpha_{\widetilde{g}t}\big)^2 \right] \\
&\geq \sum_{g=1}^{G} \widehat{\rho}^{HC} \times \min_{\widetilde{g} \in \{1, \ldots, G\}} \left[\big\| \theta^0_g - \theta_{\widetilde{g}} \big\|^2 + \frac{1}{T} \sum_{t=1}^{T} \big(\alpha^0_{gt} - \alpha_{\widetilde{g}t}\big)^2 \right],
\end{aligned}$$

where $\widehat{\rho}^{HC}$ is bounded away from zero asymptotically by Assumption S.2(a).

Moreover, using Lemma S.3, we have, as in the proof of Theorem 1,

$$(\text{S.65}) \quad \widetilde{\mathcal{Q}}\big(\widehat{\theta}^{HC}, \widehat{\alpha}^{HC}, \widehat{\gamma}^{HC}\big) - \widetilde{\mathcal{Q}}\big(\theta^0, \alpha^0, \gamma^0\big) = o_p(1).$$

[38] We use the norms (also with abuse of notation): $\|\theta_g\| = (\sum_{k=1}^{K} \theta_{gk}^2)^{1/2}$, and $\|\alpha_g\| = (\frac{1}{T} \sum_{t=1}^{T} \alpha_{gt}^2)^{1/2}$. Note that K is kept fixed as N, T tend to infinity.

Combining with (S.64), it follows that

$$\max_{g\in\{1,\dots,G\}}\left[\min_{\widetilde{g}\in\{1,\dots,G\}}\left(\left\|\theta_g^0-\widehat{\theta}_{\widetilde{g}}^{HC}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\widehat{\alpha}_{\widetilde{g}t}^{HC}\right)^2\right)\right]=o_p(1). \tag{S.66}$$

As in the proof of Lemma B.3, we then define

$$\sigma(g)=\operatorname*{argmin}_{\widetilde{g}\in\{1,\dots,G\}}\left(\left\|\theta_g^0-\widehat{\theta}_{\widetilde{g}}^{HC}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\widehat{\alpha}_{\widetilde{g}t}^{HC}\right)^2\right).$$

We have, for all $\widetilde{g}\neq g$,

$$\begin{aligned}
&\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\widehat{\theta}_{\sigma(g)}^{HC}-\widehat{\theta}_{\sigma(\widetilde{g})}^{HC}\right)+\widehat{\alpha}_{\sigma(g)t}^{HC}-\widehat{\alpha}_{\sigma(\widetilde{g})t}^{HC}\right)^2\right)^{1/2}\\
&\quad\geq\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\theta_g^0-\theta_{\widetilde{g}}^0\right)+\alpha_{gt}^0-\alpha_{\widetilde{g}t}^0\right)^2\right)^{1/2}\\
&\qquad-\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\widehat{\theta}_{\sigma(g)}^{HC}-\theta_g^0\right)+\widehat{\alpha}_{\sigma(g)t}^{HC}-\alpha_{gt}^0\right)^2\right)^{1/2}\\
&\qquad-\left(\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(x_{it}'\left(\widehat{\theta}_{\sigma(\widetilde{g})}^{HC}-\theta_{\widetilde{g}}^0\right)+\widehat{\alpha}_{\sigma(\widetilde{g})t}^{HC}-\alpha_{\widetilde{g}t}^0\right)^2\right)^{1/2}.
\end{aligned}$$

The first term on the right-hand side is asymptotically bounded away from zero by Assumption S.2(b). The second and third terms are $o_p(1)$ by (S.66). This implies that, with probability approaching 1, $\sigma(g)\neq\sigma(\widetilde{g})$. It thus follows that, with probability approaching 1, for all $\widetilde{g}\in\{1,\dots,G\}$,

$$\begin{aligned}
&\min_{g\in\{1,\dots,G\}}\left\|\theta_g^0-\widehat{\theta}_{\widetilde{g}}^{HC}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{gt}^0-\widehat{\alpha}_{\widetilde{g}t}^{HC}\right)^2\\
&\quad\leq\left\|\theta_{\sigma^{-1}(\widetilde{g})}^0-\widehat{\theta}_{\widetilde{g}}^{HC}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\sigma^{-1}(\widetilde{g})t}^0-\widehat{\alpha}_{\widetilde{g}t}^{HC}\right)^2\\
&\quad=\min_{h\in\{1,\dots,G\}}\left\|\theta_{\sigma^{-1}(\widetilde{g})}^0-\widehat{\theta}_{h}^{HC}\right\|^2+\frac{1}{T}\sum_{t=1}^{T}\left(\alpha_{\sigma^{-1}(\widetilde{g})t}^0-\widehat{\alpha}_{ht}^{HC}\right)^2,
\end{aligned}$$

which is $o_p(1)$ by (S.66).

Finally, combining with (S.66) and using the definition of the Hausdorff distance ends the proof of Lemma S.4. *Q.E.D.*

The proof of Lemma S.4 shows that there exists a permutation $\sigma:\{1,\dots,G\}\to\{1,\dots,G\}$ such that

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

By relabeling, we may take $\sigma(g)=g$. The rest of the proof of Proposition S.4 follows closely that of Theorem 2. One difference is that, in addition to bounding $\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|$, we use Assumption S.2(c) to bound $\frac{1}{T}\sum_{t=1}^{T}\|x_{it}\|^2$. The main difference with the proof of Theorem 2 comes from the fact that, instead of (B.8) one needs to show that

$$\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\left(x_{it}'\left(\theta^0_{\widetilde{g}}-\theta^0_g\right)+\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}\le-\frac{c^{HC}_{g,\widetilde{g}}}{8}\right)=o\left(T^{-\delta}\right),\tag{S.67}$$

where $c^{HC}_{g,\widetilde{g}}$ is given by Assumption S.2(b). Equation (S.67) holds because, as in the proof of Theorem 2,

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}\left(\alpha^0_{\widetilde{g}t}-\alpha^0_{gt}\right)v_{it}\right|\ge\frac{c^{HC}_{g,\widetilde{g}}}{16}\right)=o\left(T^{-\delta}\right),$$

and because, by Assumptions 1(a) and S.2(d),

$$\Pr\left(\left|\frac{1}{T}\sum_{t=1}^{T}v_{it}x_{it}'\left(\theta^0_{\widetilde{g}}-\theta^0_g\right)\right|\ge\frac{c^{HC}_{g,\widetilde{g}}}{16}\right)=o\left(T^{-\delta}\right).$$

Assumption S.2(a) in a Special Case

We consider the case where x_{it} are scalar standard normal, i.i.d. in both dimensions, and independent of g^0_j for all j. We will show that Assumption S.2(a) is satisfied. We have

$$M(\gamma,g,\widetilde{g})=\begin{pmatrix}\widehat{a} & \dfrac{1}{\sqrt{T}}\widehat{b}'\\ \dfrac{1}{\sqrt{T}}\widehat{b} & \dfrac{N(g,\widetilde{g})}{N}I_T\end{pmatrix},$$

$$\widehat{a}\sim\frac{\chi^2_{N(g,\widetilde{g})T}}{NT},\quad\text{and}\quad\widehat{b}_t\sim\frac{\mathcal{N}\left(0,N(g,\widetilde{g})\right)}{N}\quad\text{for all } t,$$

where $N(g,\widetilde{g})\equiv\sum_{i=1}^{N}\mathbf{1}\{g^0_i=g\}\mathbf{1}\{g_i=\widetilde{g}\}$.

Let $g \in \{1, \ldots, G\}$, and let $\widetilde{\delta}_g = \frac{1}{2NG} \sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\}$. For all $\theta \in \mathbb{R}$ and $\alpha \in \mathbb{R}^T$, we have

$$
\begin{aligned}
&\begin{pmatrix} \theta \\ \frac{1}{\sqrt{T}}\alpha \end{pmatrix}' M(\gamma, g, \widetilde{g}) \begin{pmatrix} \theta \\ \frac{1}{\sqrt{T}}\alpha \end{pmatrix} \\
&= \widehat{a}\theta^2 + 2\theta \frac{\alpha'\widehat{b}}{T} + \frac{N(g,\widetilde{g})}{N}\frac{\alpha'\alpha}{T} \\
&= \widehat{a}\theta^2 + 2\sqrt{\widetilde{\delta}_g}\theta \frac{\alpha'\widehat{b}}{\sqrt{\widetilde{\delta}_g}T} + \frac{N(g,\widetilde{g})}{N}\frac{\alpha'\alpha}{T} \\
&\geq \widehat{a}\theta^2 - \widetilde{\delta}_g\theta^2 - \frac{1}{\widetilde{\delta}_g}\left(\frac{\alpha'\widehat{b}}{T}\right)^2 + \frac{N(g,\widetilde{g})}{N}\frac{\alpha'\alpha}{T} \\
&\geq (\widehat{a} - \widetilde{\delta}_g)\theta^2 + \left(\frac{N(g,\widetilde{g})}{N} - \frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T}\right)\frac{\alpha'\alpha}{T},
\end{aligned}
$$

where we have used that $2ab \geq -a^2 - b^2$, and the Cauchy–Schwarz inequality.

Hence

$$
\widehat{\rho}(\gamma, g, \widetilde{g}) \geq \min\left(\widehat{a} - \widetilde{\delta}_g, \frac{N(g,\widetilde{g})}{N} - \frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T}\right).
$$

In the following derivations, we condition on $\widetilde{\delta}_g$, and omit the conditioning argument for simplicity. Using standard probability algebra, we have

$$
\begin{aligned}
&\Pr\left[\min_{\gamma\in\Gamma_G}\max_{\widetilde{g}\in\{1,\ldots,G\}} \widehat{\rho}(\gamma, g, \widetilde{g}) \leq \frac{\widetilde{\delta}_g}{2}\right] \\
&\leq \Pr\left[\min_{\gamma\in\Gamma_G}\max_{\widetilde{g}\in\{1,\ldots,G\}} \min\left(\widehat{a} - \widetilde{\delta}_g, \frac{N(g,\widetilde{g})}{N} - \frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T}\right) \leq \frac{\widetilde{\delta}_g}{2}\right] \\
&\leq G^N \max_{\gamma\in\Gamma_G} \Pr\left[\max_{\widetilde{g}\in\{1,\ldots,G\}} \min\left(\widehat{a} - \widetilde{\delta}_g, \frac{N(g,\widetilde{g})}{N} - \frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T}\right) \leq \frac{\widetilde{\delta}_g}{2}\right] \\
&\leq G^N \max_{\gamma\in\Gamma_G}\min_{\widetilde{g}\in\{1,\ldots,G\}} \Pr\left[\min\left(\widehat{a} - \widetilde{\delta}_g, \frac{N(g,\widetilde{g})}{N} - \frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T}\right) \leq \frac{\widetilde{\delta}_g}{2}\right] \\
&\leq G^N \max_{\gamma\in\Gamma_G}\min_{\widetilde{g}\in\{1,\ldots,G\}} \left(\Pr\left[\widehat{a} \leq \frac{3\widetilde{\delta}_g}{2}\right] + \Pr\left[\frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T} \geq \frac{N(g,\widetilde{g})}{N} - \frac{\widetilde{\delta}_g}{2}\right]\right),
\end{aligned}
$$

where we have used that the number of partitions $\gamma \in \Gamma_G$ is bounded by G^N.

Now, for any $(\gamma, g, \widetilde{g})$, $\widehat{a} \sim \frac{\chi^2_{N(g,\widetilde{g})T}}{NT}$ and $\frac{\widehat{b}'\widehat{b}}{T} \sim \frac{N(g,\widetilde{g})\chi^2_T}{N^2T}$. We use the following Chernoff bounds.

LEMMA S.5: *Let $Z \sim \chi^2_K$, and let $z > 0$. Then*

$$\Pr[Z \geq z] \leq \exp\left(\frac{\ln 2}{2}K - \frac{z}{4}\right),$$

$$\Pr[Z \leq z] \leq \exp\left(\frac{z}{4} - \ln\left(\frac{3}{2}\right)\frac{K}{2}\right).$$

PROOF: For all $0 < u < \frac{1}{2}$ we have, by the Markov inequality,

$$\Pr[Z \geq z] = \Pr\big[\exp(uZ) \geq \exp(uz)\big] \leq \frac{\mathbb{E}\big[\exp(uZ)\big]}{\exp(uz)} = \frac{(1-2u)^{-K/2}}{\exp(uz)},$$

where we have used the expression of the moment generating function of Z. Taking $u = \frac{1}{4}$ yields the result. The other bound is obtained similarly. *Q.E.D.*

As the groups form a partition of $\{1, \ldots, N\}$, for any γ and g there is a $\widetilde{g}$ such that

$$\frac{N(g,\widetilde{g})}{N} \geq \frac{1}{NG}\sum_{i=1}^{N} \mathbf{1}\{g_i^0 = g\} = 2\widetilde{\delta}_g.$$

For this value of $\widetilde{g}$ we have, using Lemma S.5,

$$\begin{aligned} G^N \Pr\left[\widehat{a} \leq \frac{3\widetilde{\delta}_g}{2}\right] &\leq \exp\left(N\ln(G) + \frac{3\widetilde{\delta}_g}{8}NT - \ln\left(\frac{3}{2}\right)\frac{N(g,\widetilde{g})T}{2}\right) \\ &\leq \exp\left(N\ln(G) + \frac{3\widetilde{\delta}_g}{8}NT - \ln\left(\frac{3}{2}\right)\widetilde{\delta}_g NT\right), \end{aligned}$$

where we have used that $\frac{N(g,\widetilde{g})}{N} \geq 2\widetilde{\delta}_g$.

Similarly, we have

$$\begin{aligned} G^N \Pr\left[\frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T} \geq \frac{N(g,\widetilde{g})}{N} - \frac{\widetilde{\delta}_g}{2}\right] &\leq G^N \Pr\left[\frac{\widehat{b}'\widehat{b}}{\widetilde{\delta}_g T} \geq \frac{3\widetilde{\delta}_g}{2}\right] \\ &\leq \exp\left(N\ln(G) + \frac{\ln 2}{2}T - \frac{3\widetilde{\delta}_g^2}{8}NT\right), \end{aligned}$$

where in addition we have used that $N(g,\widetilde{g}) \leq N$.

Combining results, and now indicating the conditioning of $\widetilde{\delta}_g$, we thus have

$$\begin{aligned}
&\Pr\left[\min_{\gamma\in\Gamma_G}\max_{\widetilde{g}\in\{1,\ldots,G\}}\widehat{\rho}(\gamma,g,\widetilde{g})\leq\frac{\widetilde{\delta}_g}{2}\Big|\widetilde{\delta}_g\right]\\
&\quad\leq\exp\left(N\ln(G)+\left(\frac{3}{8}-\ln\left(\frac{3}{2}\right)\right)\widetilde{\delta}_g NT\right)\\
&\qquad+\exp\left(N\ln(G)+\frac{\ln 2}{2}T-\frac{3\widetilde{\delta}_g^2}{8}NT\right).
\end{aligned}$$

Now, by Assumption 2(a), $\widetilde{\delta}_g\xrightarrow{p}\frac{\pi_g}{2G}>0$. Moreover, unconditionally,

$$\begin{aligned}
&\Pr\left[\min_{\gamma\in\Gamma_G}\max_{\widetilde{g}\in\{1,\ldots,G\}}\widehat{\rho}(\gamma,g,\widetilde{g})\leq\frac{\widetilde{\delta}_g}{2}\right]\\
&\quad\leq\Pr\left[\widetilde{\delta}_g<\frac{\pi_g}{4G}\right]+\exp\left(N\ln(G)+\left(\frac{3}{8}-\ln\left(\frac{3}{2}\right)\right)\frac{\pi_g}{4G}NT\right)\\
&\qquad+\exp\left(N\ln(G)+\frac{\ln 2}{2}T-\frac{3}{8}\left(\frac{\pi_g}{4G}\right)^2NT\right),
\end{aligned}$$

where we have used that $\frac{3}{8}<\ln(\frac{3}{2})$. As the right-hand side of this inequality tends to zero as N and T tend to infinity, this shows that

$$\min_{\gamma\in\Gamma_G}\max_{\widetilde{g}\in\{1,\ldots,G\}}\widehat{\rho}(\gamma,g,\widetilde{g})\geq\frac{\widetilde{\delta}_g}{2}+o_p(1),$$

so Assumption S.2(a) is satisfied.

REFERENCES

ACEMOGLU, D., S. JOHNSON, AND J. ROBINSON (2005): "The Rise of Europe: Atlantic Trade, Institutional Change, and Economic Growth," *American Economic Review*, 95, 546–579. [50,52]

ACEMOGLU, D., S. JOHNSON, J. ROBINSON, AND P. YARED (2008): "Income and Democracy," *American Economic Review*, 98, 808–842. [4,34,35,42,48,50,52]

ALOISE, D., P. HANSEN, AND L. LIBERTI (2012): "An Improved Column Generation Algorithm for Minimum Sum-of-Squares Clustering," *Mathematical Programming Series A*, 131, 195–220. [4-6]

ALVAREZ, J., AND M. ARELLANO (2003): "The Time Series and Cross-Section Asymptotics of Dynamic Panel Data Estimators," *Econometrica*, 71, 1121–1159. [17]

ANDERSON, T. W., AND C. HSIAO (1982): "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics*, 18, 47–82. [18]

ARCONES, M. A., AND E. GINÉ (1992): "On the Bootstrap of M-Estimators and Other Statistical Functionals," in *Exploring the Limits of Bootstrap*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, New York: Wiley, 13–47. [13]

ARELLANO, M. (1987): "Computing Robust Standard Errors for Within-Groups Estimators," *Oxford Bulletin of Economics and Statistics*, 49 (4), 431–434. [7]

ARELLANO, M., AND S. BONHOMME (2009): "Robust Priors in Nonlinear Panel Data Models," *Econometrica*, 77, 489–536. [25]

BAI, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135–171. [7]

——— (2009): "Panel Data Models With Interactive Fixed Effects," *Econometrica*, 77, 1229–1279. [14,28,29]

BAI, J., AND S. NG (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221. [14,31]

BRÜCKNER, M., AND A. CICCONE (2011): "Rain and the Democratic Window of Opportunity," *Econometrica*, 79 (3), 923–947. [53]

BRUSCO, M. J. (2006): "A Repetitive Branch-and-Bound Procedure for Minimum Within-Cluster Sums of Squares Partitioning," *Psychometrika*, 71, 347–363. [4-6]

BRUSCO, M. J., AND D. STEINLEY (2007): "A Comparison of Heuristic Procedures for Minimum Within-Cluster Sums of Squares Partitioning," *Psychometrika*, 72 (4), 583–600. [2,3,5]

CAPOROSSI, G., AND P. HANSEN (2005): "Variable Neighborhood Search for Least Squares Clusterwise Regression," Cahiers du Gerad G-005-1. [19]

CHANDA, K. C. (1974): "Strong Mixing Properties of Linear Stochastic Processes," *Journal of Applied Probability*, 11, 401–408. [21]

DU MERLE, O., P. HANSEN, B. JAUMARD, AND N. MLADENOVIC (2001): "An Interior Point Method for Minimum Sum-of-Squares Clustering," *SIAM Journal on Scientific Computing*, 21, 1485–1505. [6]

FORGY, E. W. (1965): "Cluster Analysis of Multivariate Data: Efficiency vs. Interpretability of Classifications," *Biometrics*, 21, 768–769. [1]

GINÉ, E., AND J. ZINN (1990): "Bootstrapping General Empirical Measures," *The Annals of Probability*, 18, 851–869. [13]

HANSEN, C. (2007): "Asymptotic Properties of a Robust Variance Matrix Estimator for Panel Data When T Is Large," *Journal of Econometrics*, 141 (2), 597–620. [7]

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

HANSEN, P., N. MLADENOVIĆ, AND J. A. MORENO PÉREZ (2010): "Variable Neighborhood Search: Algorithms and Applications," *Annals of Operations Research*, 175, 367–407. [2]

HUNTINGTON, S. P. (1991): *The Third Wave: Democratization in the Late Twentieth Century*. Norman, OK: University of Oklahoma Press. [33]

INABA, M., N. KATOH, AND H. IMAI (1994): "Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering," in *Proceedings of the 10th Annual Symposium on Computational Geometry*. New York: ACM Press, 332–339. [5]

KELEJIAN, H., AND I. PRUCHA (2007): "HAC Estimation in a Spatial Framework," *Journal of Econometrics*, 140, 131–154. [7]

LIN, C. C., AND S. NG (2012): "Estimation of Panel Data Models With Parameter Heterogeneity When Group Membership Is Unknown," *Journal of Econometric Methods*, 1 (1), 42–55. [19]

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

MAITRA, R., A. D. PETERSON, AND A. P. GHOSH (2011): "A Systematic Evaluation of Different Methods for Initializing the Clustering Algorithm," Unpublished Working Paper. [2]

MCLACHLAN, G., AND D. PEEL (2000): *Finite Mixture Models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. New York: Wiley-Interscience. [22]

MOON, H., AND M. WEIDNER (2010a): "Linear Regression for Panel With Unknown Number of Factors as Interactive Fixed Effects," Unpublished Manuscript. [14]

——— (2010b): "Dynamic Linear Panel Regression Models With Interactive Fixed Effects," Unpublished Manuscript. [29]

MOSCONE, F., AND E. TOSETTI (2012): "HAC Estimation in Spatial Panels," *Economics Letters*, 117, 60–65. [7]

NEWEY, W. K., AND K. D. WEST (1987): "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708. [7]

PACHECO, J., AND O. VALENCIA (2003): "Design of Hybrids for the Minimum Sum-of-Squares Clustering Problem," *Computational Statistics & Data Analysis*, 43 (2), 235–248. [2]

PAPAIOANNOU, E., AND G. SIOUROUNIS (2008): "Economic and Social Factors Driving the Third Wave of Democratization," *Journal of Comparative Economics*, 36, 365–387. [37,39,53]

POLLARD, D. (1981): "Strong Consistency of k-Means Clustering," *The Annals of Statistics*, 9, 135–140. [8,62,63]

——— (1982): "A Central Limit Theorem for k-Means Clustering," *The Annals of Probability*, 10, 919–926. [8,9,13,27,28,32,34,36,56]

POWELL, J. (1986): "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143–155. [12]

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

SPÄTH, H. (1979): "Algorithm 39: Clusterwise Linear Regression," *Computing*, 22 (4), 367–373. [19]

STEINLEY, D. (2006): "K-Means Clustering: A Half-Century Synthesis," *British Journal of Mathematical & Statistical Psychology*, 59, 1–34. [1]

Dept. of Economics, University of Chicago, 1126 E. 59th Street, Chicago, IL 60637, U.S.A.; sbonhomme@uchicago.edu

and

The Sloan School of Management, Massachusetts Institute of Technology, 100 Main Street, E62, Cambridge, MA 02142, U.S.A.; emanresa@mit.edu.

Manuscript received December, 2012; final revision received April, 2014.