Lewis, Daniel J.; Melcangi, Davide; Pilossoph, Laura; Toner-Rodgers, Aidan

Working Paper

Approximating grouped fixed effects estimation via fuzzy clustering regression

Staff Reports, No. 1033

Provided in Cooperation with:
Federal Reserve Bank of New York

Suggested Citation: Lewis, Daniel J.; Melcangi, Davide; Pilossoph, Laura; Toner-Rodgers, Aidan (2022) : Approximating grouped fixed effects estimation via fuzzy clustering regression, Staff Reports, No. 1033, Federal Reserve Bank of New York, New York, NY

This Version is available at:
https://hdl.handle.net/10419/272846

STAFF
REPORTS

NO. 1033
SEPTEMBER 2022

Approximating Grouped Fixed Effects Estimation via Fuzzy Clustering Regression

Daniel Lewis | Davide Melcangi | Laura Pilossoph | Aidan Toner-Rodgers

FEDERAL RESERVE BANK *of* NEW YORK

Approximating Grouped Fixed Effects Estimation via Fuzzy Clustering Regression

Daniel Lewis, Davide Melcangi, Laura Pilossoph, and Aidan Toner-Rodgers

Federal Reserve Bank of New York Staff Reports, no. 1033
September 2022
JEL classification: C23, C63

Abstract

We propose a new, computationally-efficient way to approximate the "grouped fixed-effects" (GFE) estimator of Bonhomme and Manresa (2015), which estimates grouped patterns of unobserved heterogeneity. To do so, we generalize the fuzzy C-means objective to regression settings. As the regularization parameter *m* approaches 1, the fuzzy clustering objective converges to the GFE objective; moreover, we recast this objective as a standard Generalized Method of Moments problem. We replicate the empirical results of Bonhomme and Manresa (2015) and show that our estimator delivers almost identical estimates. In simulations, we show that our approach delivers improvements in terms of bias, classification accuracy, and computational speed.

Key words: clustering, unobserved heterogeneity, panel data

Melcangi, Toner-Rodgers: Federal Reserve Bank of New York (emails: davide.melcangi@ny.frb.org, aidan.toner-rodgers@ny.frb.org). Lewis: University College London (email: daniel.j.lewis@ucl.ac.uk). Pilossoph: Duke University (email: laura.pilossoph@duke.edu).

 This research was supported in part through computational resources provided by the BigTex High Performance Computing Group at the Federal Reserve Bank of Dallas.

1 Introduction

In a recent paper, Bonhomme and Manresa (2015) (henceforth BM) propose the "grouped fixed-effects estimator," a form of K-means regression, to study grouped patterns of heterogeneity in panel data settings. In particular, the authors study the linear model:

$$y_i = \sum_{g=1}^{G} \gamma_{ig} \theta_g x_i + \nu_i, \tag{1}$$

where $y_i \in \mathbb{R}^T$, $x_i \in \mathbb{R}^K$, (x_i, y_i) are independently distributed across i (and identically distributed conditional on group membership, $\tilde{g}_i$), $\gamma_{ig} = \mathbf{1}[\tilde{g}_i = g]$ and $E[\nu_i | x_i, \tilde{g}_i = g] = 0$. θ_g is a $T \times K$ matrix of group-specific coefficients on the covariates x, for $g = 1, \ldots, G$.[1] The indicators γ_{ig} are equal to 1 if observation i belongs to group g, and zero otherwise. Equation (1) postulates that the outcomes y are generated linearly from x, with parameters depending on observation i's group membership, which is unobservable to the econometrician. BM are concerned with estimating θ_g for all G groups in addition to group membership indicators $\gamma_{ig} = \mathbf{1}[i \in g]$, where x_i may simply be a constant term, hence the "grouped fixed-effects" (GFE) estimator.

In this paper, we provide a method to approximate the GFE objective function that incorporates a degree of regularization of the membership function, allowing for computationally more efficient estimation and delivering more precise estimates. Our approach is particularly valuable with large data sets, which are increasingly the norm in applied economics, as well as when the specified number of groups G becomes large.

BM consider the population GFE least-squares criterion

$$(\tilde{\theta}, \tilde{\gamma}) = \arg\min_{\theta, \gamma} E\left[\sum_{g=1}^{G} \gamma_{ig} \left\| y_i - \theta_g x_i \right\|^2\right], \tag{2}$$

where the minimum is taken over all possible group membership assignments, $\tilde{g} = \{\tilde{g}_1, ..., \tilde{g}_N\}$ of the N observed entities, and the group-specific effects (θ_g), which collectively form θ. Instead, we study

$$\left(\tilde{\theta}_m^{FCR}, \tilde{\mu}\right) = \arg\min_{\theta, \mu} \left[\sum_{g=1}^{G} \mu_{ig}^m \left\| y_i - \theta_g x_i \right\|^2\right], \tag{3}$$

where the minimum is taken over the weights μ_{ig} and the group-specific effects θ_g, given some user-specified regularization parameter $m > 1$. Mean clustering problems of the form in Equation (3) are known as "fuzzy C-means" algorithms, due to the fact that $m > 1$ induces continuous weights and thus "fuzzy" group assignments, rather than the binary assignments of "hard" K-means as in Equation (2). We thus refer to the model in Equation (3) as "fuzzy C-regression," or

[1] As we will discuss later, this notation is general enough to allow for common coefficients both across groups and time.

FCR.

We show that in the limit as $m \to 1$, the "fuzzy" objective in Equation (3) is equivalent to the K-means objective in Equation (2). However, by writing the optimal weights μ_{ig} as a function of m and θ, FCR can be reframed as a Generalized Method of Moments (GMM) problem. It is then easy to show that for a fixed m, a sample estimator $\hat{\theta}_m^{FCR}$ is consistent for $\tilde{\theta}_m^{FCR}$, with an asymptotically normal limiting distribution.

Why is this a useful representation of the original problem in Equation (2)? First, as noted above, as m approaches 1, the "fuzzy" objective function converges to the original GFE objective function, providing a convenient approximation. We find in both simulations and BM's empirical application that for $m = 1.001$, say, the quality of the approximation is excellent, as evidenced by the performance of $\hat{\theta}_m^{FCR}$ relative to either the true parameters and group assignment, or BM's empirical estimates. Second, because the objective in Equation (3) is continuous for $m > 1$ (but potentially close to 1), it is natural to estimate $\tilde{\theta}_m^{FCR}$ not through an iterative procedure, alternating between choosing a set of discrete group membership functions and conditionally optimizing the model parameters, as is conventional for clustering problems, but rather through direct minimization of the objective function, as in any other GMM problem. This means the problem can be solved in a single step using any built-in minimization routine, without the need to explicitly code the specialized algorithms described in BM. Indeed, the preferred solution method of BM involves searching over essentially all possible group assignments of the N entities, which is a very computationally intensive task as N and G grow. Perhaps due to resulting numerical challenges of searching over the binary group assignments, we find that the ability of our approximating estimator to recover the truth (in terms of both parameter estimates and group assignment) is superior to the original GFE estimator, in spite of the lesser computation time noted above.

In summary, we find that in simulations our approximating estimator yields more reliable parameter estimates, lower misclassification rates, and far shorter computation time, particularly as $N, T,$ and G grow. We thus think of the fuzzy clustering approach as a fast and easy way to estimate grouped fixed-effects.

The paper proceeds as follows. In Section 2, we formally introduce the fuzzy clustering problem and show that the objective converges to the GFE objective as $m \to 1$ and that it can be recast as a standard GMM problem. Section 3 replicates the results of BM's income and democracy empirical application, reports the analogous estimates using FCR, and compares the two estimators in an extension of BM's simulation exercise. Section 4 concludes.

2 Fuzzy Clustering Regression

In this section we describe the fuzzy clustering regression (FCR) objective and methodology. We first show that it converges to the "hard" K-means objective of BM, justifying the use of m close

to 1 to approximate GFE estimators. Next, we argue that the FCR problem can be rewritten as a standard GMM problem, providing asymptotic results.

2.1 FCR approximates GFE

Consider the linear model introduced in Equation (2),

$$y_i = \sum_{g=1}^{G} \gamma_{ig} \theta_g x_i + \nu_i, \tag{4}$$

where $y_i \in \mathbb{R}^T$, $x_i \in \mathbb{R}^K$, (x_i, y_i) are independently distributed across i (and identically distributed conditional on group membership, $\tilde{g}_i$), $\gamma_{ig} = \mathbf{1}[\tilde{g}_i = g]$ and $E[\nu_i | x_i, \tilde{g}_i = g] = 0$. θ_g is a $T \times K$ matrix representing the group-specific coefficients on the covariates x. The $t-$dimension allows for a panel structure, with repeated observations of each entity i over time, or simply multiple observed outcomes for each entity. The "hard" K-means (HKM), or in the terminology of BM, GFE, objective function is:

$$J^{HKM}(\theta) = E\left[\min_g \left\|y - \theta_g x\right\|^2\right], \tag{5}$$

Alternatively, the objective above can be rewritten as a weighted sum replacing minimization with the resulting binary group membership function,

$$L^{HKM}(\theta) = E\left[\sum_{g=1}^{G} \gamma_g^*(y, x; \theta) \left\|y - \theta_g x\right\|^2\right], \tag{6}$$

where $\gamma_g^*(y, x; \theta) = \mathbf{1}\left[\left\|y - \theta_g x\right\|^2 \leq \|y - \theta_h x\|^2 \ \forall\, h \neq g\right]$.

The FCR objective function introduces regularization for the estimated group membership function (previously $\gamma_g^*(\cdot)$), so that it is no longer binary. In particular, the objective function is instead

$$L_m^{FCR}(\theta, \mu) = E\left[\sum_{g=1}^{G} \mu_g^m \left\|y - \theta_g x\right\|^2\right] \tag{7}$$

where $m > 1$ is the regularization parameter and μ_g represent group weights. Bezdek (1981) derives the MSE optimal weights μ_g, given m and θ, $\mu_g(y, x; \theta, m)$, in the cluster means case (i.e., a single constant regressor in x). Analogously, the optimal weights in the present regression case are found by taking the derivative of Equation (7) with respect to μ_g subject to the constraint that the weights $\mu_g(y, x; \theta, m)$ sum to unity:

$$\mu_g(y, x; \theta, m) = \left(\sum_{h=1}^{G} \frac{\left\|y - \theta_g x\right\|^{2/(m-1)}}{\left\|y - \theta_h x\right\|^{2/(m-1)}}\right)^{-1}, g = 1, \ldots, G, \tag{8}$$

Yang and Yu (1992) show that in the cluster means case the objective in Equation (7) can be rewritten subsuming the optimal parameter-dependent weights in Equation (8) into the objective function itself. For fixed m, define $\mu(y, x; \theta, m) = (\mu_1(y, x; \theta, m), \ldots, \mu_G(y, x; \theta, m))$. Then, Equation (7) can be rewritten as

$$\begin{aligned} L_m^{FCR}(\theta, \mu) &= L_m^{FCR}(\theta, \mu(y, x; \theta, m)) \\ &= E\left[\sum_{g=1}^{G}\left(\sum_{h=1}^{G}\frac{\|y-\theta_g x\|^{2/(m-1)}}{\|y-\theta_h x\|^{2/(m-1)}}\right)^{-m}\|y-\theta_g x\|^2\right] \\ &= E\left[\left(\sum_{g=1}^{G}\|y-\theta_g x\|^{-2/(m-1)}\right)^{1-m}\right] \\ &\equiv J_m^{FCR}(\theta) \end{aligned} \tag{9}$$

Significantly, $J_m^{FCR}(\theta)$ has replaced an objective function with weights linked to the parameters with a nonlinear function in (the norm of) the group-specific errors $\|y-\theta_g x\|$, since the weights themselves are simply a function of those errors.

We introduce standard conditions in Assumption 1 to argue that $J_m^{FCR}(\theta)$ approximates $J^{HKM}(\theta)$.

Assumption 1. $E\left[\|y\|^2\right] < \infty$, $E\left[\|x\|^2\right] < \infty$, *and* $\theta \in \Theta$, *which is compact.*

Proposition 1. *Under Assumption 1,*

$$\lim_{m \to 1^+} J_m^{FCR}(\theta) = J^{HKM}(\theta),$$

Proposition 1 shows that the FCR objective function converges to the HKM (or GFE) objective function as $m \to 1$ from above. Thus, setting the regularization parameter close to 1 allows the econometrician to approximate the GFE clustering problem using the FCR objective function. However, for $m > 1$, she retains the continuity of the FCR membership function, ensured by the regularization, and the benefits we describe below.

It is natural to ask what conclusions can be drawn about the minimizers of the FCR objective function. In the limit, it is true that

$$\tilde{\theta}_1^{FCR} = \arg\min \lim_{m \to 1^+} J_m^{FCR}(\theta) = \tilde{\theta}. \tag{10}$$

However, since the convergence in Proposition 1 is not uniform in θ, y, x, it is not in general the case that $\lim_{m \to 1^+} \tilde{\theta}_m^{FCR} = \tilde{\theta}$, so we view $\tilde{\theta}_m^{FCR}$ as a regularized approximation to $\tilde{\theta}$, as opposed to a convergent estimand. Nevertheless, we are able to characterize the limiting distribution of $\tilde{\theta}_m^{FCR}$ for fixed m, which we consider below.

2.2 FCR as GMM

The FCR objective in Equation (9) is differentiable for fixed m, and the first order conditions, stated in Proposition 2, constitute a set of just-identifying moment conditions, as described in Proposition 2.

Proposition 2. *The solution $\tilde{\theta}_m^{FCR}$ satisfies the moment conditions*

$$E\left[\left(\sum_{h=1}^{G}\frac{\left\|y_i-\theta_g x_i\right\|^{2/(m-1)}}{\left\|y_i-\theta_h x_i\right\|^{2/(m-1)}}\right)^{-m}\left(y_{it}-\theta_{g(t)}x_i\right)x_i\right]=0 \text{ for } g=1,\ldots,G \text{ and } t=1,\ldots,T, \quad (11)$$

where t indexes dimensions of y_i and (t) rows of θ_g; FCR is a GMM problem.

This result has two main implications. First, $\tilde{\theta}_m^{FCR}$ can be estimated via direct minimization of the FCR objective function. This is a standard non-linear minimization problem. Typically, clustering problems are solved iteratively, alternating between assigning entities to the best-fitting group (or computing their fuzzy or probabilistic weights for each group), then re-estimating model parameters. The econometrician is required to search over many possible group assignments, particularly in the case of binary weights, as in the GFE estimator. Such algorithms, where individual observations or subsets of observations are systematically reallocated between groups from one iteration to the next, are prone to local minima. Unfortunately, they are not well-suited to parallelization, since the search over possible groupings is usually conducted sequentially. While the direct minimization required by FCR may also be prone to local minima, it is straightforward to parallelize across start values to mitigate such concerns, and the computation is fast enough to facilitate many start values.

Second, the asymptotic properties of both HKM (or GFE) and FCM problems have previously proven difficult to establish, requiring extensive technical arguments (e.g., Pollard (1981, 1982); Yang and Yu (1992); Yang (1994); Bonhomme and Manresa (2015)). However, Proposition 2 shows that the FCR clustering problem is simply a GMM problem, and the asymptotic properties of the estimator (for fixed m) follow by standard arguments.

We now present the asymptotic properties of the sample analog of $\tilde{\theta}_m^{FCR}$. Define

$$S_N(\theta)=\frac{1}{N}\sum_{i=1}^{N}\eta\left(\theta,y_i,x_i\right)'\sum_{i=1}^{N}\eta\left(\theta,y_i,x_i\right), \quad (12)$$

where the $(G\times T\times K)\times 1$ vector-valued moment function $\eta\left(\theta,y_i,x_i\right)$ stacks

$$\left[\left(\sum_{h=1}^{G}\frac{\left\|y_i-\theta_g x_i\right\|^{2/(m-1)}}{\left\|y_i-\theta_h x_i\right\|^{2/(m-1)}}\right)^{-m}\left(y_{it}-\theta_{g(t)}x_i\right)x_i\right]=0$$

for $g = 1, \ldots, G$ and $t = 1, \ldots, T$. The FCR estimator is

$$\hat{\theta}_m^{FCR} = \arg\min_{\theta} S_N(\theta). \tag{13}$$

Assumption 2 presents standard assumptions for the GMM estimator's asymptotic properties, which are presented in Proposition 3. We present assumptions and results in a large-N, fixed-T framework, which we believe is a better representation of typical datasets, particularly in macroeconomic settings, than a large-N, T framework.

Assumption 2. *(*Consistency and asymptotic normality*)*

1. *The observations $i = 1, \ldots, N$ are independently (and identically, within groups) sampled,*
2. *G is finite,*
3. *$\tilde{\theta}_m^{FCR}$ is the unique solution to $E\left[\eta\left(\theta, y_i, x_i\right)\right] = 0$ (up to ordering of the groups).*
4. *$\tilde{\theta}_m^{FCR}$ is in the interior of Θ,*
5. *$H = E\left[\frac{\partial \eta(\theta, y_i, x_i)}{\partial \theta'}\right]$ is full rank,*
6. *$E\left[\sup_{\theta \in \mathcal{N}} \left\| \frac{\partial \eta(\theta, y_i, x_i)}{\partial \theta'} \right\|\right] < \infty$ in a neighborhood $\mathcal{N}$ of $\tilde{\theta}_m^{FCR}$,*
7. *$V = E\left[\eta\left(\tilde{\theta}_m^{FCR}, y_{i,}, x_i\right) \eta\left(\tilde{\theta}_m^{FCR}, y_{i,}, x_i\right)'\right]$ is positive definite.*

Assumption 2.1 assumes that the entities are sampled independently, but makes no assumption about the dependence properties of possible repeated observations of each entity. While conditional on group membership observations come from different distributions (since at least θ_g varies), viewing group membership as a latent variable, the data is unconditionally i.i.d. Assumption 2.2 ensures that the number of groups does not increase at the same rate as the sample size. Assumption 2.3 stipulates that the population solution to the FCR problem is unique. For nonlinear problems like this, primitive conditions for the identification assumption are challenging to provide, but in the limit (as $m \to 1^+$), the identification condition becomes the same as that of GFE, which simply requires that for a given group, θ_g is identified in the regression $y_i = \theta_g x_i + v_i, \tilde{g}_i = g$. The remaining conditions are standard technical assumptions.

Proposition 3.

1. *Under Assumptions 1 and 2.1-2.3, $\hat{\theta}_m^{FCR} \xrightarrow{p} \tilde{\theta}_m^{FCR}$,*
2. *Under Assumptions 1 and 2, $\sqrt{N}\left(\hat{\theta}_m^{FCR} - \tilde{\theta}_m^{FCR}\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1} V H^{-1}\right)$.*

Proposition 3 shows that for fixed m, the estimator $\hat{\theta}_m^{FCR}$ is consistent for $\tilde{\theta}_m^{FCR}$, the population minimizer, with an asymptotically normal limiting distribution. As discussed above, in the limit (as $m \to 1^+$), $\tilde{\theta}_m^{FCR}$ is equal to $\tilde{\theta}$. We provide expressions for the Hessian in Appendix C.

As previously noted, we consider large-N, fixed-T asymptotics, since we believe they are a better representation of the data structure and uncertainty in typical datasets of interest. However, with straightforward modifications to the assumptions (e.g., weak dependence over T), it is possible to also establish large-T limiting distributions with convergence at a faster $\sqrt{NT}$ rate. In contrast, BM focus on the large-N, T framework, with analogous fixed-T results in their appendix. Importantly, under the fixed-T framework, estimated group membership is subject to uncertainty, while under large-T it can be recovered with certainty, assuming weak dependence. Therefore, the fixed-T distributions that we present are conservative relative to large-T distributions.

Additionally, the moment conditions in Proposition 2 can easily accommodate regressors with common coefficients across groups, $\theta_{gtk} = \theta_{htk}$, or across dimensions of y_i or periods, $\theta_{gtk} = \theta_{hsk}$.[2] In the former case, it is straightforward to show that the corresponding moment condition is

$$E\left[\left(\sum_{h=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2/(m-1)}\right)^{-m}\sum_{g=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2m/(m-1)}\left(y_{it}-\theta_{g(t)}x_i\right)x_{ik}\right]=0$$

and in the latter,

$$E\left[\left(\sum_{h=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2/(m-1)}\right)^{-m}\sum_{t=1}^{T}\sum_{g=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2m/(m-1)}\left(y_{it}-\theta_{g(t)}x_i\right)x_{ik}\right]=0.$$

To summarize, we have argued that for suitably chosen m close to 1, the FCR objective function approximates the GFE objective function, with equality in the limit. The FCR problem facilitates a GMM implementation with direct minimization, unlike the iterative procedures used to implement most clustering algorithms. For fixed m, the corresponding finite sample estimator is consistent for its population counterpart, which, in the limit, is equal to the GFE solution. The estimator is also asymptotically normal, and analytical standard errors are available for the parameter estimates.

In this paper, we focus on FCR as a tool to approximate GFE. However, the results above hold for any $m > 1$. Completely separately, FCR with larger m may also be a valuable tool in its own right in many economic applications. Indeed, "fuzzy C-means" algorithms, which FCR generalizes, were introduced as a form of "possibilistic" clustering that could better accommodate the uncertainty of group membership in realistic datasets, where noise means that cluster membership cannot be ascertained with certainty. Setting a higher value of the regularization parameter

[2]For example, in the forthcoming application, the parameter set θ encompasses both group-specific time-varying constants $\alpha_{g(t)}$, as well as group- and time-invariant coefficients on the common covariates $democracy_{it-1}$ and $logGDPpc_{it-1}$.

m flattens the group membership function, increasing the inherent uncertainty of the assignment. Thus, FCR can also be used as a non-parametric alternative to finite mixture models, such as the Gaussian Mixture Model considered in Lewis, Melcangi and Pilossoph (2022).

3 Application and Simulations

We first replicate the empirical results of BM using their estimator and replication code. Then, to illustrate our approach, we implement the FCR estimator using the same specifications, and extend BM's simulations calibrated to that application to include our proposed estimator. Doing so allows us to directly compare the performance of the two methods in terms of accuracy, efficiency, and computational speed. In particular, following BM, we consider the panel of countries from Acemoglu, Johnson, Robinson and Yared (2008), who study the coevolution of income and democracy from 1970–2000.[3] Our specification takes the following form, where we regress an index of democracy (the Freedom House indicator) on its lagged value, lagged log GDP per capita, and the group-time effect α_{gt}:

$$democracy_{it} = \beta_1 democracy_{it-1} + \beta_2 logGDPpc_{it-1} + \alpha_{g_i t} + \nu_{it}, \tag{14}$$

where g_i denotes the group membership of country i.

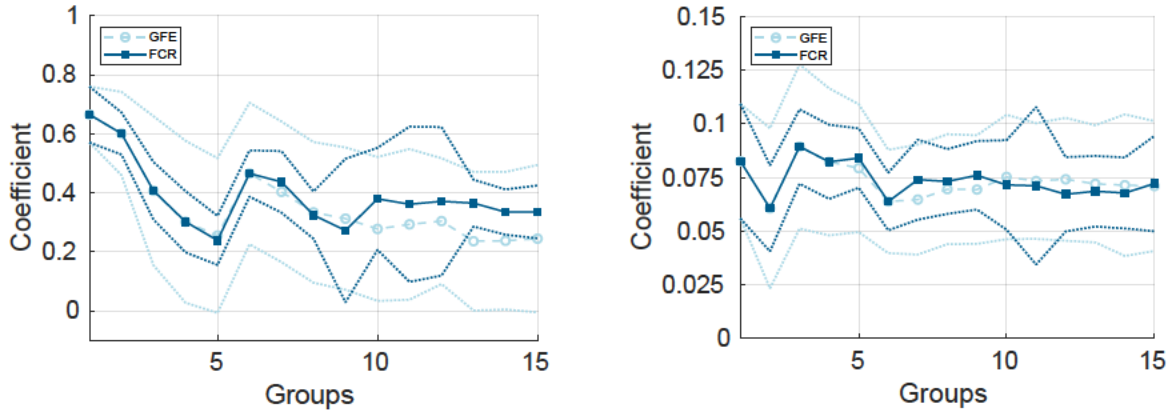
3.1 Computational Details

As discussed in Section 2, the "fuzziness" of the FCR membership function is governed by the regularization parameter m, where group assignment becomes binary as $m \to 1^+$. After some experimentation, we set $m = 1.001$ in order to replicate the BM estimates. In practice, this generates group weights equal to either 0 or 1 to 6 decimal places. In our main results, we supply the algorithm with 1,000 starting values and select the set of objective-minimizing coefficients across iterations.[4]

To mitigate concerns over possible local minima, we run FCR using multiple starting values. Table B1 shows the performance of our estimator for a varying number of starting values; in our main results we choose 1,000 starting values, which we find offers the best mix of performance and computational speed. Importantly, our algorithm can be parallelized across starting values, and our main estimation is run with 250 parallel cores.

[3] We use the balanced panel from Acemoglu et al. (2008), which includes 90 countries observed over seven five-year periods. All data files are available for download at: http://economics.mit.edu/files/5000.

[4] We have also estimated the model using 100 (Appendix B) and 10,000 (unreported) starting values.



(a) Democracy (β_1) (b) Income (β_2)

Figure 1: Coefficients for Lagged Democracy and Income

Notes: This figure plots FCR and GFE estimates for the lagged democracy and lagged income coefficients using the Acemoglu et al. (2008) balanced panel. The horizontal axis indicates the number of groups, while the vertical axis shows parameter values. The dotted lines show 95% confidence intervals for the two sets of estimates. The GFE estimates and confidence bands are estimated using BM's replication code and match Figure 1 in BM, where the standard errors are calculated using their preferred bootstrap method. For FCR, we construct the confidence bands using analytic asymptotic standard errors.

3.2 Replication of BM Results

First, we replicate the empirical results reported in BM using their estimator and replication code. These are plotted using the dashed line with circles in Figure 1. Next, we show that our approach is able to closely replicate these empirical results of BM. Figure 1 plots the estimated coefficients on lagged democracy and lagged income for $G = 1$ to $G = 15$ from both algorithms, along with their respective 95% confidence intervals.[5] The two methods produce nearly identical point estimates, especially for small G, and on average and in absolute terms the results differ by 0.049 for β_1 and 0.002 for β_2.[6] Our analytical standard errors tend to be somewhat narrower than BM's bootstrap intervals, since while both account for the finite nature of $T = 7$, the bootstrap intervals additionally account for the finite nature of N. However, it is important to note that bootstrapping in this context introduces substantial additional computational burden, as the model needs to be re-estimated for each bootstrapped sample. Next, we plot the time-varying group-specific effects for $G = 4$, shown in Figure 2. Again, we are able to closely match the coefficients of BM; the estimates are identical to at least 4 decimal places and the mean absolute difference is 0.000008.

[5]The BM standard errors are calculated using their preferred bootstrap method, while the FCR standard errors are analytical asymptotic standard errors, which is why they are tighter than BM's.

[6]As we show in the next section, in simulation FCR performs better on measures of bias, suggesting that in cases where FCR and GFE slighlty diverge, the FCR estimate may in fact be closer to the truth, on average.

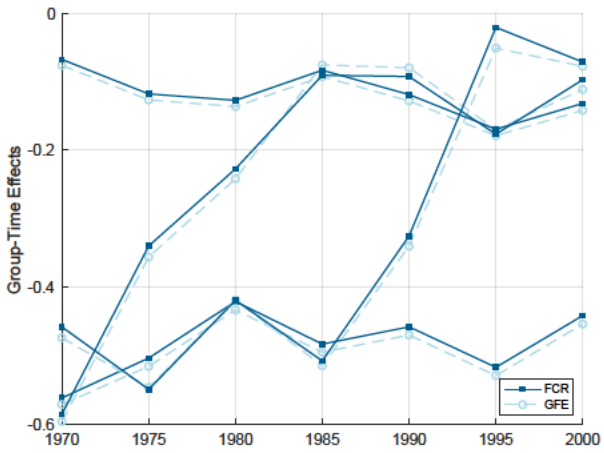


Figure 2: Group-Time Effects with $G = 4$

Notes: This figure plots FCR group-time effect estimates using the Acemoglu et al. (2008) balanced panel with four groups. The x-axis indicates the year from 1970-2000, while the y-axis plots parameter values. The solid lines show FCR results and the dotted lines represent GFE estimates (shown in Figure 2 in BM).

3.3 Simulations and FCR Performance

After showing that FCR is a convenient approximation to the GFE estimator, we now compare the performance of the two methods through simulation. We start from the simulation code found in BM's replication package, and extend it to include our estimator. In particular, we simulate a panel of the same size as the application ($N = 90$, $T = 7$), where data is generated by first estimating the GFE model on the empirical dataset and then using the resulting group-time, common coefficient, and binary group membership estimates to create the DGP. In particular,

$$y_{it}^{s} = \hat{\alpha}_{g_i t} + \hat{\beta}_1 democracy_{it-1} + \hat{\beta}_2 logGDPpc_{it-1} + \nu_{it}^{s}$$

where $\hat{g}_i$ is the estimated group membership from BM, $\hat{\alpha}_{g_i t}$ is the estimated group-time effect, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated common coefficients, and the errors ν_{it}^{s} are i.i.d. normal draws for simulation sample s, with variance equal to the mean squared residual. We consider 1000 Monte Carlo samples.

The simulation results are presented in Table 1, where we show the performance of the two estimators in terms of bias, group misclassification rate, inference, and computation time. FCR compares favorably on all measures, particularly misclassification when G is large. When $G = 10$, for example, the GFE estimator classifies 44.7 percent of units incorrectly, while FCR misclassifies only 16.11 percent. Across simulations, computation time is similar for the two estimators, with GFE slightly faster for some specifications given the particular choice of tuning parameters. However, note that the tuning paramters used in BM's simulation are not their generally

Table 1: Simulation Performance

	$G = 3$		$G = 5$		$G = 10$	
	FCR	GFE	FCR	GFE	FCR	GFE
Bias						
Average lagged democracy bias	0.035	0.084	0.042	0.056	0.051	0.054
Average lagged democracy RMSE	0.043	0.094	0.056	0.070	0.067	0.075
Average lagged income bias	0.013	0.032	0.010	0.007	0.009	0.013
Average lagged income RMSE	0.016	0.035	0.012	0.017	0.012	0.015
Group Misclassification						
Average misclassification rate	9.37%	9.50%	7.69%	9.68%	16.11%	44.73%
Inference						
Median lagged democracy standard errors	0.051	0.051	0.050	0.068	0.056	0.048
Coverage rate for lagged democracy	0.894	0.790	0.911	0.840	0.937	0.940
Median lagged income standard errors	0.011	0.013	0.010	0.014	0.014	0.010
Coverage rate for lagged income	0.885	0.840	0.938	0.960	0.939	0.930
Computation Time						
Total time (seconds)	17.5	24.8	27.6	26.5	127.4	78.2

Notes: This table compares the performance of the FCR and GFE estimators on a simulated panel with $N = 90$ and $T = 7$. GFE results come from Tables S3 and S4 in their paper. While BM estimate the model with their preferred specification (Algorithm 2, with 10 starting values, 10 neighbors, and maximum steps of 10), they use only 5 starting values and 5 maximum steps in simulations, since their baseline specification "resulted in prohibitive computation times". For consistency within this table, we compute computation times ourselves under this specification. Table S1 and S2 in BM report computation times for their baseline algorithm: 38.4 and 228.4 seconds for 3 and 10 groups, respectively. The GFE standard errors and coverage use the Pollard (1982) fixed-T formula, reported in columns (2) of BM Table S7. Bias, misclassification, computation time, and non-rejection probability are means across 1,000 simulations, while the standard errors are medians (to match the reported estimates in BM). For FCR, we use 1,000 starting values and 250 parallel cores; the reported computation time is the total across these starting values, not the time per starting value.

recommended tuning parameters, compared which FCR proves universally faster.[7] In terms of inference, the median standard errors for the common coefficients, a measure reported by BM, are closely comparable across the two estimators. On the other hand, the coverage rate of the 95% confidence intervals is generally much closer to its nominal level using the analytical standard errors we derive than those proposed by BM under the fixed-T asymptotic framework.

The fuzzy clustering approximation is particularly valuable in settings with large panels, which are increasingly common in empirical work. Specifically, the regularization incorporated in the FCR objective permits direct minimization in a single step, which is substantially faster than previously implemented approaches, even after parallelization across many start values. To illustrate

[7] In line with BM, GFE simulation results are presented with a faster Algorithm 2 (see note to the table) than their preferred specification. BM reports that computation time for their preferred algorithm is 38.4 and 228.4 seconds for 3 and 10 groups, respectively, thus slower than FCR. We also present a more thorough analysis of computation time in the next exercise, revealing that FCR is considerably faster on larger datasets).

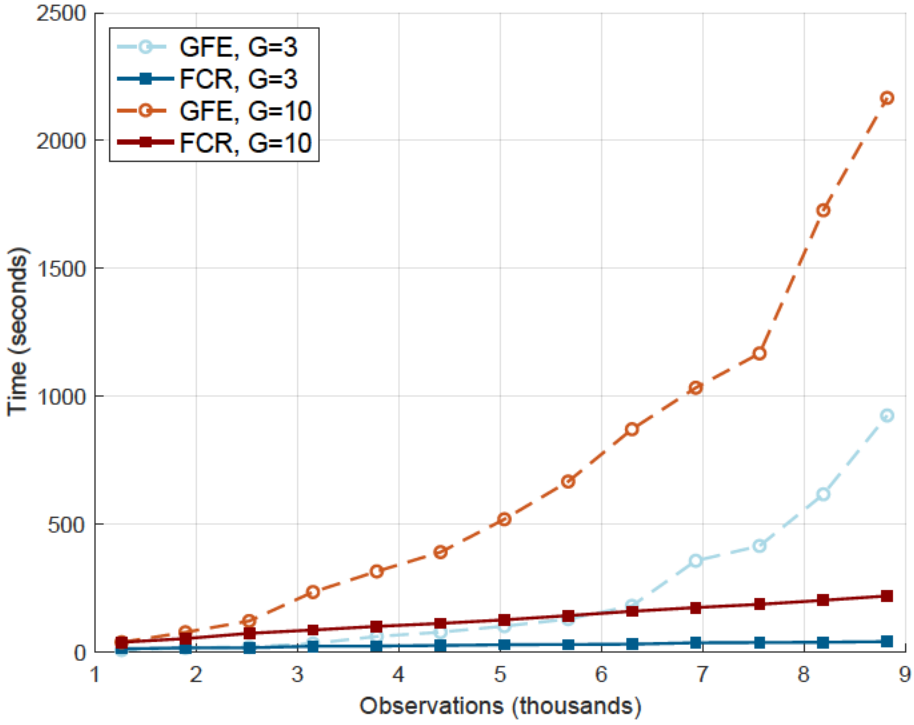


Figure 3: Computation Time by Dataset Size

Notes: This figure plots computation time by dataset size for the FCR and GFE estimators. The horizontal axis represents number of observations in the dataset ($N * T$) and the vertical axis plots computation time in seconds. The solid blue line shows FCR with 3 groups, the solid red line shows FCR with 10 groups, the dashed blue line shows GFE with 3 groups, and the dashed orange line shows GFE with 10 groups. The datasets are scaled in the N dimension (with T fixed at 7) and we construct these larger datasets by stacking the simulated data from our first exercise, with independent error draws for each observation. The BM GFE estimator is run using their "Algorithm 2" with 10 simulations, 10 neighbors, and maximum steps of 10, their preferred specification used in their main estimation exercises. Because their code is designed with a manual step for entering options, this is necessarily included in the computation time for GFE; however, this accounted for a negligible share of the overall computational time (approximately 10 seconds). The FCR algorithm is run with 1,000 starting values and 250 parallel cores.

this point, we plot computation time by dataset size in Figure 3, where the data is expanded in the N dimension. We construct these larger datasets by stacking the simulated panels from our first exercise, with independent error draws for each observation. While FCR remains quite efficient and appears to scale linearly with number of observations, computation time for the GFE estimator becomes prohibitive even on moderately sized datasets and grows nonlinearly. Thus, we see fuzzy clustering as a fast and easy way to estimate grouped patterns of heterogeneity, with particular advantages in larger datasets.

4 Conclusion

In this paper, we have proposed a new, computationally efficient way to approximate the "grouped fixed-effects" estimator of BM, which estimates grouped patterns of unobserved heterogeneity. To

do so, we generalized the fuzzy C-means objective to regression settings, and showed that as the regularization parameter m approaches 1, the fuzzy clustering objective converges to the GFE objective. The fuzzy clustering formulation allows us to recast the problem as a standard GMM problem, instead of an iterative group assignment and minimization problem, and to derive a standard limiting distribution. We replicate the empirical results of BM using their estimator, and show that our approach produces very similar estimates. In simulations, we show that our estimator exhibits a smaller bias than previously suggested approaches and achieves substantially more accurate classification of individual observations as the number of groups increases. Moreover, our approach delivers a dramatic reduction in computation time as the sample size increases.

References

Acemoglu, Daron, Simon Johnson, James A. Robinson, and Pierre Yared, "Income and Democracy," *American Economic Review*, 2008, *98* (3), 808–42.

Bezdek, James, *Pattern Recognition With Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

Bonhomme, Stéphane and Elena Manresa, "Grouped Patterns of Heterogeneity in Panel Data," *Econometrica*, 2015, *83* (3), 1147–1184.

Hayashi, Fumio, *Econometrics*, Princeton University Press, 2011.

Lewis, Daniel J., Davide Melcangi, and Laura Pilossoph, "Latent heterogeneity in the marginal propensity to consume," Staff Report 902, New York, NY 2022.

Newey, Whitney K. and Daniel McFadden, "Large Sample Estimation and Hypothesis Testing," in "in," Vol. 4 of *Handbook of Econometrics*, Elsevier, 1994, pp. 2111 – 2245.

Pollard, David, "Strong Consistency of K-Means Clustering," *The Annals of Statistics*, 1981, *9* (1), 135–140.

_ , "A Central Limit Theorem for K-Means Clustering," *The Annals of Probability*, 1982, *10* (4), 919–926.

Spivak, Michael, *Calculus On Manifolds: A Modern Approach To Classical Theorems Of Advanced Calculus*, Avalon Publishing, 1971.

Yang, Miin-Shen, "On Asymptotic Normality of a Class of Fuzzy C-Means Clustering Procedures," *International Journal of General Systems*, 1994, *22* (4), 391–403.

_ **and Kai Fun Yu**, "On Existence and Strong Consistency of a Class of Fuzzy C-Means Clustering Procedures," *Cybernetics and Systems*, 1992, *23* (6), 583–602.

A Proofs

Proof of Proposition 1

Proof. Consider the limit of the intermediate formulation of the FCR objective,

$$\lim_{m\to 1^+} E\left[\sum_{g=1}^{G}\left(\sum_{h=1}^{G}\frac{\left\|y-\theta_g x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}}\right)^{-m}\left\|y-\theta_g x\right\|^2\right]. \tag{15}$$

First, we can move the limit inside the expectation by the dominated convergence theorem; this follows since for all y, x, θ, $\mu(y, x; \theta, m)$, the only part of the integrand that depends on m, is bounded above by 1 (and below by zero), and $E\left[\left\|y-\theta_g x\right\|^2\right]$ is finite by Assumption 1.

We next take logs and evaluate the limit for a given group, g, and arbitrary y, x, θ:

$$\lim_{m\to 1^+}\log\left(\sum_{h=1}^{G}\frac{\left\|y-\theta_g x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}}\right)^{-m}\left\|y-\theta_g x\right\|^2 \tag{16}$$

$$=\lim_{m\to 1^+} -m\log\sum_{h=1}^{G}\frac{\left\|y-\theta_g x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}}+\log\left\|y-\theta_g x\right\|^2. \tag{17}$$

To conserve space, denote $\epsilon_g = y - \theta_g x$. We first focus on the weights, working term by term in the summation over h. There are three cases to consider:

$$\lim_{m\to 1^+}\frac{\left\|\epsilon_g\right\|^{2/(m-1)}}{\left\|\epsilon_h\right\|^{2/(m-1)}}=\begin{cases}\infty, & \|\epsilon_h\|<\|\epsilon_g\|\\ 0, & \|\epsilon_h\|>\|\epsilon_g\|\\ 1, & \|\epsilon_h\|=\|\epsilon_g\|.\end{cases} \tag{18}$$

Putting together the summation over h, these intermediate limits imply the following three cases:

$$\lim_{m\to 1^+}\sum_{h=1}^{G}\frac{\left\|\epsilon_g\right\|^{2/(m-1)}}{\left\|\epsilon_h\right\|^{2/(m-1)}}=\begin{cases}1, & \|\epsilon_g\|<\|\epsilon_h\|\ \forall h\neq g\\ \infty, & \exists h:\|\epsilon_g\|>\|\epsilon_h\|\\ n^*, & \text{otherwise,}\end{cases} \tag{19}$$

where the final case applies when $n^* > 1$ group assignments yield the identical minimum norm residual, $\|\epsilon\|^*$.

Then, using the fact that the limit of a log equals the log of the limit, we have

$$\lim_{m\to 1^+} -m \log \sum_{h=1}^{G} \frac{\left\|\epsilon_g\right\|^{2/(m-1)}}{\left\|\epsilon_h\right\|^{2/(m-1)}} = \begin{cases} 0, & \left\|\epsilon_g\right\| < \left\|\epsilon_h\right\| \forall h \neq g \\ -\infty, & \exists h : \left\|\epsilon_g\right\| > \left\|\epsilon_h\right\| \\ -\log n^*, & \text{otherwise.} \end{cases} \tag{20}$$

Finally, exponentiating, since the log of the desired limit equals the limit of the logs we have computed, yields

$$\lim_{m\to 1^+} \left(\sum_{h=1}^{G} \frac{\left\|\epsilon_g\right\|^{2/(m-1)}}{\left\|\epsilon_h\right\|^{2/(m-1)}} \right)^{-m} \left\|\epsilon_g\right\|^2 = \begin{cases} \left\|\epsilon_g\right\|^2, & \left\|\epsilon_g\right\| < \left\|\epsilon_h\right\| \forall h \neq g \\ 0, & \exists h : \left\|\epsilon_g\right\| > \left\|\epsilon_h\right\| \\ 1/n^* \left\|\epsilon_g\right\|^2, & \text{otherwise.} \end{cases} \tag{21}$$

Therefore, in the first two cases, each term in the outer summation over g converges to their GFE counterparts, with binary indicators equal to 1 if $\left\|\epsilon_g\right\|$ is a unique minimum, and zero if it is larger than the minimum, so the summation itself converges to that for GFE, for any y, x, θ. In the final, knife-edge, case, where multiple assignments yield the same minimum residual, GFE and other HKM estimators typically impose a fixed membership rule (i.e., set the indicator equal to 1 for the group with the lowest index for which the minimum value of the objective is obtained, and zero for the others) in order to prevent the algorithm from perpetually permuting the membership amongst equivalent groups. However, while the limit of the summand for each g may not equal that under the arbitrary decision rule of GFE in this case, the limit of the summation over g is still identical to that of GFE, since for all such groups, j, $\left\| \epsilon_j\right\| = \left\| \epsilon\right\|^*$, and

$$\lim_{m\to 1^+} \left(\sum_{h=1}^{G} \frac{\left\|\epsilon_g\right\|^{2/(m-1)}}{\left\|\epsilon_h\right\|^{2/(m-1)}} \right)^{-m} \left\|\epsilon_g\right\|^2 = n^* \times \frac{1}{n^*} \left\|\epsilon\right\|^{*2} = \left\|\epsilon\right\|^{*2},$$

as for GFE. Finally, since we argued above that limit and expectation were interchangeable, we have

$$\lim_{m\to 1^+} E\left[\sum_{g=1}^{G} \left(\sum_{h=1}^{G} \frac{\left\|y-\theta_g x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}} \right)^{-m} \left\|y-\theta_g x\right\|^2 \right] = \lim_{m\to 1^+} J_m^{FCR}(\theta) = J^{HKM}(\theta) \tag{22}$$

□

Proof of Proposition 2

Proof. We start by differentiating the argument of the expectation in $J_m^{FCR}(\theta)$ with respect to θ_{gtk}:

$$
\begin{aligned}
&\frac{\partial}{\partial \theta_{gtk}} \left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{1-m} \\
&= (1-m) \left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{-m} \frac{\partial}{\partial \theta_{gtk}} \sum_{g=1}^{G} \left\| y - \theta_g x \right\|^{-2/(m-1)} \\
&= (1-m) \left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{-m} \frac{-2}{m-1} \left\| y - \theta_g x \right\|^{(1+m)/(1-m)} \frac{\partial}{\partial \theta_{gtk}} \left\| y - \theta_g x \right\| \\
&= 2 \left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{-m} \left\| y - \theta_g x \right\|^{-(1+m)/(m-1)} \frac{y_t - \theta_{g(t)} x}{\left\| y - \theta_g x \right\|} (-x_k) \\
&= -2 \left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{-m} \left\| y - \theta_g x \right\|^{-2m/(m-1)} \left(y_t - \theta_{g(t)} x \right) x_k \\
&= -2 \left(\sum_{h=1}^{G} \frac{\left\| y - \theta_g x \right\|^{2/(m-1)}}{\left\| y - \theta_h x \right\|^{2/(m-1)}} \right)^{-m} \left(y_t - \theta_{g(t)} x \right) x_k,
\end{aligned}
$$

where $\theta_{g(t)}$ denotes the row of θ_g corresponding to outcome y_t. Note that since these partial derivatives are continuous in θ (by inspection; see also Yang (1994) Lemma 2), the function is (continuously) differentiable in θ (Spivak (1971) Theorem 2.8). Moreover, $\left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{1-m}$ is Lebesgue-integrable for each θ as

$$
\left(\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2/(m-1)} \right)^{1-m} \leq \sum_{h=1}^{G} \left\| y - \theta_h x \right\|^{-2(1-m)/(m-1)} = \sum_{h=1}^{G} \left\| y - \theta_h x \right\|^2
$$

since $1 - m < 0$ and

$$
\begin{aligned}
\sum_{h=1}^{G} \left\| y - \theta_h x \right\|^2 &\leq \sum_{h=1}^{G} \left(\|y\| + \|\theta_h x\| \right)^2 \\
&\leq \sum_{h=1}^{G} \left(\|y\| + \|\theta_h\| \, \|x\| \right)^2 \\
&= \sum_{h=1}^{G} \|y\|^2 + 2 \|\theta_h\| \, \|x\| \, \|y\| + \|\theta_h\|^2 \|x\|^2, \qquad (23)
\end{aligned}
$$

which is integrable by Assumption 1. Moreover, (23) establishes a bounding function for the integrand in terms of θ. From these conditions, the dominated convergence theorem allows the interchange of differentiation and integration:

$$\begin{aligned}
\frac{\partial J_m^{FCR}(\theta)}{\partial \theta_{gtk}} &= \frac{\partial}{\partial \theta_{gtk}} E\left[\left(\sum_{g=1}^{G}\left\|y-\theta_g x\right\|^{-2/(m-1)}\right)^{1-m}\right] \\
&= E\left[\left(\frac{\partial}{\partial \theta_{gtk}}\sum_{g=1}^{G}\left\|y-\theta_g x\right\|^{-2/(m-1)}\right)^{1-m}\right] \\
&= E\left[\left(\frac{\partial}{\partial \theta_{gtk}}\sum_{g=1}^{G}\left\|y_i-\theta_g x_i\right\|^{-2/(m-1)}\right)^{1-m}\right] \\
&= E\left[-2\left(\sum_{h=1}^{G}\frac{\left\|y_i-\theta_g x_i\right\|^{2/(m-1)}}{\left\|y_i-\theta_h x_i\right\|^{2/(m-1)}}\right)^{-m}\left(y_{it}-\theta_{g(t)}x_i\right)x_{ik}\right].
\end{aligned}$$

Stacking the conditions vertically for row t of θ_g yields the $k \times 1$ vector

$$\frac{\partial J_m^{FCR}}{\partial \theta'_{g(t)}} = E\left[-2\left(\sum_{h=1}^{G}\frac{\left\|y-\theta_g x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}}\right)^{-m}\left(y_t-\theta_{g(t)}x\right)x\right].$$

Proceeding likewise across $t = 1, \ldots, T$ and for $g = 1, \ldots, G$ yields $G \times T \times k$ conditions which θ^* must satisfy,

$$E\left[\left(\sum_{h=1}^{G}\frac{\left\|y_i-\theta_g x_i\right\|^{2/(m-1)}}{\left\|y_i-\theta_h x_i\right\|^{2/(m-1)}}\right)^{-m}\left(y_{it}-\theta_{g(t)}x_i\right)x_i\right] = 0, \text{ for } g = 1, \ldots, G, t = 1, \ldots, T,$$

since θ^* minimizes $J_m^{FCR}(\theta)$. These $G \times T \times K$ equations constitute continuous moment conditions for the $G \times T \times K$ free parameters in θ. Thus, the system of equations constitutes a just-identified GMM problem. □

Proof of Proposition 3.1

Proof. By Assumption 2.1, (y_i, x_i) are independently (and identically, given group membership) distributed. Viewing group membership, $\tilde{g}_i$, as a latent state variable that is drawn alongside the observed variables and innovations (and of which y_i and potentially x_i and v_i are functions), the full vector of observed and unobserved variables is unconditionally i.i.d. (although the distribution of y_i is clearly group-dependent, conditional on the state variable $\tilde{g}_i$). By Assumption 2.3, θ^* uniquely satisfies $\eta(\theta, y_i, x_i)$. As noted in the proof of Proposition 2, the moment conditions $\eta(\theta, y_i, x_i)$ are continuous for all $\theta \in \Theta$, so the objective function is also. Next, we show that the moments are bounded in expectation for all $\theta \in \Theta$ (the dominance condition). Observe that $\left(\sum_{h=1}^{G}\frac{\left\|y-\theta_g x\right\|^{2/(m-1)}}{\left\|y-\theta_h x\right\|^{2/(m-1)}}\right)^{-m}$ is bounded between zero and one (the supremum of the summation is

infinity as the residuals $y - \theta_h x$, $h \neq g$ go to zero and the infimum is 1 as $y - \theta_h x$, $h \neq g$ go to infinity). So

$$\begin{aligned} E\left[\sup_{\theta\in\Theta}\|\eta\left(\theta, y_i, x_i\right)\|\right] &\leq E\left[\sup_{\theta\in\Theta}\sup_{g}\left\|\left(y_i - \theta_g x_i\right)x_i'\right\|\right] \\ &= E\left[\sup_{\theta\in\Theta}\sup_{g}\left\|y_i x_i' - \theta_g x_i x_i'\right\|\right] \\ &\leq E\left[\sup_{\theta\in\Theta}\sup_{g}\left\|y_i x_i'\right\| + \left\|\theta_g x_i x_i'\right\|\right] \\ &\leq E\left[\sup_{\theta\in\Theta}\sup_{g}\|y_i\|\,\|x_i\| + \left\|\theta_g\right\|\,\|x_i\|\,\|x_i\|\right] \\ &< \infty, \end{aligned}$$

where the third inequality follows from the triangle inequality, the fourth from Cauchy-Schwarz, and the final follows from Assumption 1. These points jointly satisfy the requirements of standard GMM arguments, (e.g., Newey and McFadden (1994), p. 2121—2, Hayashi (2011) Proposition 7.7), so $\hat{\theta}_m^{FCR} \xrightarrow{p} \tilde{\theta}_m^{FCR}$. □

Proof of Proposition 3.2

Proof. First, we provide expressions for H to establish the continuous differentiability of $\eta\left(\theta, y_i, x_i\right)$ in θ. We focus on the cross-sectional case here ($T = 1$) for the sake of simplicity, but provide expressions for panel data corresponding to the setting of our empirical application and simulation study in Section C.2. Partition the blocks of H as

$$H = \begin{bmatrix} H_{11} & \cdots & H_{1g} & \cdots & H_{1G} \\ \vdots & \ddots & & & \vdots \\ H_{g1} & & H_{gg} & & H_{gG} \\ \vdots & & & \ddots & \vdots \\ H_{G1} & \cdots & H_{Gg} & \cdots & H_{GG} \end{bmatrix},$$

where $H_{gh} = \frac{\partial^2 J_m^{FCR}}{\partial\theta_g \partial\theta_h'}$ with $H_{gh} = H_{hg}'$ by symmetry of the Hessian. For the case where all coefficients are group-specific, it can be shown that

$$\begin{aligned} H_{gg} &= E\left[x_i x_i'\left\{\frac{-2m}{m-1}A_i^{-m-1}\left(e_{gi}\right)^2 C_{gi}^2 + \frac{m+1}{m-1}A_i^{-m}C_{gi}\right\}\right] \\ H_{gh} &= E\left[x_i x_i'\left\{\frac{-2m}{m-1}A_i^{-m-1}C_{hi}e_{hi}e_{gi}C_{gi}\right\}\right], h \neq g, \end{aligned}$$

where $e_{gi} = y_i - \theta_g x_i$, $A_i = \Sigma_{g=1}^G \left\|e_{gi}\right\|^{-2/(m-1)}$, $C_{gi} = \left\|e_{gi}\right\|^{-2m/(m-1)}$. We also provide expressions for additional elements of the Hessian when there are covariates with common coefficients across groups, such that $\theta_{gk} = \theta_{hk} \equiv \theta_{\star k}, h \neq g$. In this case,

$$\begin{aligned}
\frac{\partial^2 J_m^{FCR}}{\partial\theta_{\star k}\partial\theta_{\star k}} &E\left[x_{ik}^2\left\{\frac{-2m}{m-1}A_i^{-m-1}B_i^2 + \frac{m+1}{m-1}A_i^{-m}\sum_{g=1}^G C_{gi}\right\}\right] \\
\frac{\partial^2 J_m^{FCR}}{\partial\theta_{\star k}\partial\theta_{\star l}} &= E\left[x_{ik}x_{i,l}\left\{\frac{-2m}{m-1}A_i^{-m-1}B_i^2 + \frac{m+1}{m-1}A_i^{-m}\sum_{g=1}^G C_{gi}\right\}\right] \\
\frac{\partial^2 J_m^{FCR}}{\partial\theta_{\star k}\partial\theta_{gl}} &= E\left[x_{ik}x_{i,l}\left\{\frac{-2m}{m-1}A_i^{-m-1}C_{gi}e_{gi}B_i + \frac{m+1}{m-1}A_i^{-m}C_{gi}\right\}\right],
\end{aligned}$$

where $B_i = \sum_{g=1}^G \left[e_{gi}C_{gi}\right]$. By inspection, all elements of these Hessians are continuous in θ, since e_{gi}, A_i^{-m}, A_i^{-m-1}, C_{gi}, B_i are continuous in θ, and all elements of H are continuous functions of these objects.

Next, the asymptotic normality of $\frac{1}{\sqrt{N}}\sum_{i=1}^N \eta\left(\theta, y_i, x_i\right)$ follows by Lindeberg-Lévy Central Limit Theorem by Assumptions 1 and 2.1,

$$\frac{1}{\sqrt{N}}\sum_{i=1}^N \eta\left(\theta^*, y_i, x_i\right) \xrightarrow{d} \mathcal{N}\left(0, V\right),$$

where $V = E\left[\eta\left(\theta, y_i, x_i\right)\eta\left(\theta, y_i, x_i\right)'\right]$ is assumed to be positive definite in Assumption 2.7.

Combining these two results with the remaining conditions of Assumption 2, the standard conditions for asymptotic normality of a GMM estimator are satisfied (e.g., Hayashi (2011) Proposition 7.10). Since the weighting matrix is the identity (the problem is just-identified),

$$\sqrt{N}\left(\hat{\theta}_m^{FCR} - \tilde{\theta}_m^{FCR}\right) \xrightarrow{d} \mathcal{N}\left(0, H^{-1}VH^{-1}\right).$$

□

B Additional Results

Table B1: FCR Performance for Varying Number of Starting Values

	$G = 3$					$G = 5$					$G = 10$				
Starting Values	1	10	50	100	1,000	1	10	50	100	1,000	1	10	50	100	1,000
Bias															
Average lagged democracy bias	0.129	0.039	0.035	0.035	0.035	0.114	0.117	0.111	0.078	0.042	0.128	0.116	0.114	0.114	0.051
Average lagged democracy RMSE	0.254	0.112	0.098	0.055	0.094	0.381	0.224	0.129	0.089	0.056	0.255	0.166	0.148	0.141	0.067
Average lagged income bias	0.014	0.014	0.013	0.013	0.013	0.018	0.019	0.017	0.017	0.010	0.016	0.016	0.015	0.015	0.009
Average lagged income RMSE	0.033	0.034	0.022	0.021	0.016	0.038	0.034	0.027	0.025	0.012	0.039	0.038	0.034	0.022	0.012
Group Misclassification															
Average misclassification rate	14.95%	11.31%	9.37%	9.37%	9.37%	18.44%	17.87%	14.12%	9.31%	7.69%	48.25%	34.64%	25.77%	18.36%	16.11%
Computation Time															
Total time (seconds)	1.9	5.7	9.3	12.1	17.5	2.67	12.3	18.1	21.2	27.6	12.34	28.1	58.5	78.5	127.4

Notes: This table shows FCR performance and computation time for 1, 10, 50, 100, and 1,000 starting values. The results for 1 starting value are run with a single core, while the rest are run with the full 250 parallel workers for best comparison (the overhead time could be further optimized by using a tailored number of parallel workers depending on the number of starting values).

C Gradient and Hessian of Objective Function

In this section, we provide explicit expressions for the gradient and Hessian of the objective function, $L_m^{FCR}(\theta)$, which is equivalent to the moment function exploited by the GMM estimator, specialized to the case considered in our empirical application and simulation study, with time-varying group-specific intercepts, α_{gt}, and other coefficients constant across groups and time. Let s index such additional common covariates, and denote the corresponding coefficients as ψ_s. For the purposes of this discussion, instead of denoting repeated observations of some regressor s for observation i as different generic elements k of the vector x_i, we instead refer to them directly by the it subscript (e.g., x_{its}), to be consistent with BM.

C.1 Gradient

Define $e_{git} = y_{it} - \alpha_{gt} - \sum_{s=1}^{S} \theta_s x_{its}$. The sample gradient of the objective with respect to α_{gt} is

$$\frac{\partial \hat{L}_m^{FCR}(\theta)}{\partial \alpha_{gt}} = -2 \sum_{i=1}^{N} \left[\sum_{h=1}^{G} \frac{\left\{ \sum_t \left(e_{git} \right)^2 \right\}^{\frac{1}{m-1}}}{\left\{ \sum_t \left(e_{hit} \right)^2 \right\}^{\frac{1}{m-1}}} \right]^{-m} e_{git}.$$

For a common coefficient, ψ_s, the gradient is

$$\frac{\partial \hat{L}_m^{FCR}(\theta)}{\partial \psi_s} = -2 \sum_{i=1}^{N} \left[\sum_{h=1}^{G} \frac{1}{\left\{ \sum_{t=1}^{T} \left(e_{hit} \right)^2 \right\}^{\frac{1}{m-1}}} \right]^{-m} \sum_{g=1}^{G} \left\{ \left\{ \sum_{t=1}^{T} \left(e_{git} \right)^2 \right\}^{\frac{-m}{m-1}} \sum_{t=1}^{T} e_{git} x_{its} \right\}$$

Define $a_i = \sum_{g=1}^{G} \frac{1}{\left\{ \sum_{t=1}^{T} \left(e_{git} \right)^2 \right\}^{\frac{1}{m-1}}}$. Then $\hat{L}_m^{FCR}(\theta) = \sum_{i=1}^{N} a_i^{1-m}$. Then simplifying the notation, the gradient for a group-specific time-varying fixed effect is

$$D_{\alpha_{gt}} = \frac{\partial \hat{L}_m^{FCR}(\theta)}{\partial \alpha_{gt}} = -2 \sum_{i=1}^{N} a_i^{-m} \left\{ \sum_{t=1}^{T} \left(e_{git} \right)^2 \right\}^{\frac{-m}{m-1}} e_{git}$$

and the gradient for a common coefficient is

$$D_{\psi_s} = \frac{\partial \hat{L}_m^{FCR}(\psi_s)}{\partial \theta_s} = -2 \sum_{i=1}^{N} a_i^{-m} \sum_{g=1}^{G} \left\{ \left\{ \sum_{t=1}^{T} \left(e_{git} \right)^2 \right\}^{\frac{-m}{m-1}} \sum_{t=1}^{T} e_{git} x_{its} \right\}$$

$$\nabla_L = \left[D'_{\alpha_{11}} \dots D'_{\alpha_{1T}} \;\; D'_{\alpha_{21}} \dots D'_{\alpha_{2T}} \dots D'_{\alpha_{gt}} \;\; D'_{\psi_1} \dots D'_{\psi_S} \right]'$$

C.2 Hessian

We now provide explicit expressions for the Hessian corresponding to our empirical application and simulation study. We start by partitioning the Hessian as

$$H = \begin{bmatrix}
H_{\alpha_{11},\alpha_{11}} & \cdots & H_{\alpha_{11},\alpha_{1T}} \cdots & H_{\alpha_{11},\alpha_{G1}} \cdots & H_{\alpha_{11},\alpha_{GT}} \cdots & H_{\alpha_{11},\psi_1} \cdots & H_{\alpha_{11},\psi_S} \\
\vdots & \ddots & & & & & \vdots \\
H_{\alpha_{1T},\alpha_{11}} & \cdots & H_{\alpha_{1T},\alpha_{1T}} \cdots & H_{\alpha_{1T},\alpha_{G1}} \cdots & H_{\alpha_{1T},\alpha_{GT}} \cdots & H_{\alpha_{1T},\psi_1} \cdots & H_{\alpha_{1T},\psi_S} \\
\vdots & & \ddots & & & & \vdots \\
H_{\alpha_{G1},\alpha_{11}} & \cdots & H_{\alpha_{G1},\alpha_{1T}} \cdots & H_{\alpha_{G1},\alpha_{G1}} \cdots & H_{\alpha_{G1},\alpha_{GT}} \cdots & H_{\alpha_{G1},\psi_1} \cdots & H_{\alpha_{G1},\psi_S} \\
\vdots & & & \ddots & & & \vdots \\
H_{\alpha_{GT},\alpha_{11}} & \cdots & H_{\alpha_{GT},\alpha_{1T}} \cdots & H_{\alpha_{GT},\alpha_{G1}} \cdots & H_{\alpha_{GT},\alpha_{GT}} \cdots & H_{\alpha_{GT},\psi_1} \cdots & H_{\alpha_{GT},\psi_S} \\
\vdots & \ddots & & & & & \vdots \\
H_{\psi_1,\alpha_{11}} & \cdots & H_{\psi_1,\alpha_{1T}} \cdots & H_{\psi_1,\alpha_{G1}} \cdots & H_{\psi_1,\alpha_{GT}} \cdots & H_{\psi_1,\psi_1} \cdots & H_{\psi_1,\psi_S} \\
\vdots & \ddots & & & & & \vdots \\
H_{\psi_S,\alpha_{11}} & \cdots & H_{\psi_S,\alpha_{1T}} \cdots & H_{\psi_S,\alpha_{G1}} \cdots & H_{\psi_S,\alpha_{GT}} \cdots & H_{\psi_S,\psi_1} \cdots & H_{\psi_S,\psi_S}
\end{bmatrix}$$

where $H_{\alpha_{11},\alpha_{1t}} = \frac{\partial D_{\alpha_{11}}}{\partial \alpha_{1t}}$, and so on.

We first characterize the diagonal elements, corresponding to the group-specific intercepts, which take the form

$$\begin{aligned}
H_{\alpha_{gt},\alpha_{gt}} = 4 \sum_{i=1}^{N} \frac{m}{m-1} a_i^{-m-1} \left(\sum_{t=1}^{T} \left(e_{git} \right)^2 \right)^{\frac{-2m}{m-1}} e_{git}^2 \\
+4 \sum_{i=1}^{N} a_i^{-m} \frac{-m}{m-1} \left\{ \sum_{t=1}^{T} \left(e_{git} \right)^2 \right\}^{\frac{-m}{m-1}-1} e_{git}^2 \\
+2 \sum_{i=1}^{N} a_i^{-m} \left\{ \sum_{t=1}^{T} \left(e_{git} \right)^2 \right\}^{\frac{-m}{m-1}}.
\end{aligned}$$

Next, entries corresponding to intercepts for the same group, g, but different time periods, t are given by

$$H_{\alpha_{gt},\alpha_{gt'}} = 4\sum_{i=1}^{N}\frac{m}{m-1}a_i^{-m-1}\left(\sum_{t=1}^{T}\left(e_{git}\right)^2\right)^{\frac{-2m}{m-1}}\left(e_{git'}\right)e_{git} \\ +4\sum_{i=1}^{N}a_i^{-m}\frac{-m}{m-1}\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}-1}e_{git'}e_{git}$$

and conversely, for the same t, but different g,

$$H_{\alpha_{gt},\alpha_{g't}} = 4\sum_{i=1}^{N}\frac{m}{m-1}a_i^{-m-1}\left(\sum_{t=1}^{T}\left(e_{g'it}\right)^2\right)^{\frac{-1}{m-1}-1}\left(e_{g'it}\right)\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}}e_{git}.$$

Entries corresponding to a pair of intercepts from different groups and time periods are given by

$$H_{\alpha_{gt},\alpha_{g't'}} = 4\sum_{i=1}^{N}\frac{m}{m-1}a_i^{-m-1}\left(\sum_{t=1}^{T}\left(e_{g'it}\right)^2\right)^{\frac{-1}{m-1}-1}\left(e_{g'it'}\right)\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}}e_{git}.$$

Entries corresponding to a group-specific intercept and a common coefficient have the form

$$H_{\alpha_{gt},\psi_s} = \frac{4m}{m-1}\sum_{i=1}^{N}a_i^{-m-1}\left\{\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-1}{m-1}-1}\sum_{t=1}^{T}\left(e_{git}x_{its}\right)\right\}\right\}\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}}e_{git} \\ -\frac{4m}{m-1}\sum_{i=1}^{N}a_i^{-m}\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}-1}\left\{\sum_{t=1}^{T}\left(e_{git}x_{its}\right)\right\}e_{git} \\ +2\sum_{i=1}^{N}a_i^{-m}\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}}\left(x_{its}\right).$$

Finally, diagonal entries corresponding to common coefficients are

$$H_{\psi_s,\psi_s} = \frac{4m}{m-1}\sum_{i=1}^{N}a_i^{-m-1}\left[\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-1}{m-1}-1}\sum_{t=1}^{T}\left(e_{git}x_{its}\right)\right\}\right]^2 \\ -\frac{4m}{m-1}\sum_{i=1}^{N}a_i^{-m}\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}-1}\left(\sum_{t=1}^{T}e_{git}x_{its}\right)^2\right\} \\ -2\sum_{i=1}^{N}a_i^{-m}\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}}\sum_{t=1}^{T}\left[-x_{its}x_{its}+e_{git}\right]\right\}$$

and off-diagonal entries corresponding to a pair of two different common coefficients are given by

$$
\begin{aligned}
H_{\psi_s,\psi_{s'}} = \frac{4m}{m-1}\sum_{i=1}^{N} a_i^{-m-1}\left[\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-1}{m-1}-1}\sum_{t=1}^{T}\left(\left(e_{git}x_{its'}\right)\right)\right\}\right]^2 \\
-\frac{4m}{m-1}\sum_{i=1}^{N} a_i^{-m}\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}-1}\left(\sum_{t=1}^{T} e_{git}x_{its'}\right)\sum_{t=1}^{T} e_{git}x_{its}\right\} \\
+2\sum_{i=1}^{N} a_i^{-m}\sum_{g=1}^{G}\left\{\left\{\sum_{t=1}^{T}\left(e_{git}\right)^2\right\}^{\frac{-m}{m-1}}\sum_{t=1}^{T}\left[x_{its}x_{its'}\right]\right\}
\end{aligned}
$$