Replication package for "Approximating Grouped Fixed Effects Estimation via Fuzzy Clustering Regression" by Daniel Lewis, Davide Melcangi, Laura Pilossoph, and Aidan Toner-Rodgers

Required Matlab Toolboxes

Our code is run using Matlab 2022A and requires the following toolboxes:

- Statistics and machine learning toolbox
- Optimization toolbox
- Parallel computing toolbox

How to Run

The bash file `runall.sh` runs all the necessary Matlab files to reproduce the tables and figures in the main text and appendix.

- It first runs `fig1.m` and and `fig2.m`, which apply our estimator to the data of Bonhomme and Manresa (2015) (henceforth BM) and plot a comparison of our estimates with grouped fixed effects.
- Next, it runs `simulate_panel.m`, which sets up the simulated data for our subsequent exercises.
- Using this simulated data, we then run `table1.m` and `tableB1.m`, which calls our estimator using a variety of starting values and group numbers to produce the results for Table 1 in the main text and Table B1 in the appendix.
- Finally, `fig3.m` runs our estimator on a number of dataset sizes to produce Figure 3.

Repository Structure

- `data/raw`: contains the raw data, which come from the BM replication files
- `data/intermediate`: stores intermediate files
- `code`: contains file to produce all tables and figures in the main text and appendix, using functions stored in `code/functions`
- `output`: stores results for all tables and figures

Description of Data

All our data come from BM. Specifically, we use `final_data.mat` which is the dataset used in their empirical application. `BM_LHS_panel.mat` and `BM_RHS_panel.mat` simply split this dataset into the outcome and covariates of our regression specification, respectively. Additionally, we use the files `BM_coeffs.mat` and `BM_fe_4G.mat` which are coefficient estimates stored in the BM replication package (and replicated by us).

For further details on the data including variable definitions see the BM replication package.

Parallelization

Our main estimation is run with 250 parallel cores. However, the code can be run with any number of cores (just adjust `parpool`) although this will change computation time.

Bonhomme and Manresa (2015) Replication

The replication of the Bonhomme and Manresa (2015) results is run using their replication code, which can be found here.