Code Documentation

Group Heterogeneity in Panel Data Models – A Test to Determine the Unknown Number of Groups

Caroline Krayer

1 Overview

This documentation contains an overview and brief description of the different R files contained in the Github repository of my master thesis presented to the Department of Economics at Rheinische-Friedrich-Wilhelms-Universität Bonn. The repository is set up as an RStudio project (*MA-repo.Rproj*) that needs to be opened for reproducibility.

In my thesis, I deal with group heterogeneous panel data models and study a testing procedure developed by Lu and Su (2017) that aims to determine the unknown number of groups in the population. In the applied part I design a simulation study to investigate the weaknesses and strengths of the approach and employ the method to a real data example studying the effect of democracy on economic growth (see Acemoglu et al., 2019).

2 R-Packages Required

An important part of my code builds upon the R-package *classo* that implements the group heterogeneous estimator of Su, Shi and Phillips (2016) for models without time-fixed effects. The beta version needs to be installed from Github as specified here. It also requires the installation of *Rtools* and *Rmosek*. The installation of *Rmosek* is described here or here. Importantly, *Rmosek* needs an academic license that can be obtained from this website.

Further R-packages that need to be installed to run my code are:

cowplot	*matrixStats*	*reshape2*	*tis*
devtools	*parallel*	*rlist*	*tmap*
dplyr	*pbapply*	*sf*	*xtable*
expm	*plm*	*SparseM*	
ggplot2	*plyr*	*spData*	
MASS	*pracma*	*stats*	
Matrix	*readstata13*	*tidyverse*	

3 Overview and Description of the R-Files

Functions

(a) *bias_correction.R* contains a function that employs half-panel jackknife to correct for the asymptotic bias in the group-specific parameter estimates of the data application.

(b) *calculate_residuals.R* computes the regression residuals of a pre-specified model.

(c) *demean_data.R* contains functions that eliminate the individual- and time-fixed effects in panel data models as well as a function for models with time-fixed effects that stacks the regressors in a matrix needed in C-Lasso estimation.

(d) *DGP.R* contains the implementation of the data generating processes used in the simulation study.

(e) *estimators.R* implements various estimators that I use as a comparison. Among them are the fixed-effects estimator for homogenous panel data models, the completely heterogeneous coefficient estimator, the infeasible estimator for group heterogeneous models, and Swamy's random coefficient estimator.

(f) *lasso_est_time.R* implements C-Lasso for models with time-fixed effects.

(g) *se_classo.R* computes the analytical standard errors of the group-specific C-Lasso estimates.

(h) *test_stat.R* contains an implementation of the Lagrange-multiplier type hypothesis test developed by Lu and Su (2017).

Simstudy

(a) *comparison_estimators.R* computes the mean squared error, bias and empirical standard errors across the Monte Carlo repetitions for various estimators. It can be used to replicate Table A.4 in Appendix A.

(b) *distribution_plot.R* plots the distribution of the test statistic across the Monte Carlo repetitions and can be used to produce Figure 5.1.

(c) *frequencies.R* computes the empirical rejection frequencies of the hypothesis tests as well as the estimated number of groups with the testing procedure and information criterion, respectively. It can be used to replicate the results of Tables A.1 to A.3 in Appendix A.

(d) *powerfunction.R* computes the power of the hypothesis test when deviating from parameter homogeneity and increasing the deviation.

(e) *powerfunction_plot.R* plots the power of the hypothesis test when deviating from parameter homogeneity and thus replicates Figure 5.2.

(f) *simstudy.R* simulates the testing procedure, the information criterion and the parameter estimates.

Application

(a) The folder *Data* contains the data for the empirical application – Acemoglu et al. (2019)'s original data *DDCGdata_final.dta*, and the data for the balanced subpanel *democracy-balanced-l4.dta* on which I rely (see Chen, Chernozhukov and Fernández-Val, 2019).

(b) *analysis_data_application_democracy.R* further analyses the estimated groups and produces the results reported in Section 6.2.

(c) *data_application.R* implements the testing procedure and parameter estimation for a general data application.

(d) *data_application_democracy.R* performs the data application in Section 6, analysing the effect of democracy on economic growth.

(e) *tables_data_application_democracy.R* tables the results of the testing procedure, the information criterion and the parameter estimates of the empirical application in Section 6.1.

(f) *worldmap.R* maps the group membership of each country included in the data application and thus reproduces Figure 6.1.

Bibliography

Acemoglu, D., S. Naidu, P. Restrepo and J. A Robinson (2019). "Democracy Does Cause Growth". In: *Journal of Political Economy*, Vol. 127, No. 1, pp. 47–100.

Chen, S., V. Chernozhukov and I. Fernández-Val (2019). "Mastering Panel Metrics: Causal Impact of Democracy on Growth". In: *AEA Papers and Proceedings*. Vol. 109, pp. 77–82.

Lu, X. and L. Su (2017). "Determining the Number of Groups in Latent Panel Structures with an Application to Income and Democracy." In: *Quantitative Economics*, Vol. 8, No. 3, pp. 729–760.

Su, L., Z. Shi and P. C. B. Phillips (2016). "Identifying Latent Structures in Panel Data". In: *Econometrica*, Vol. 84, No. 6, pp. 2215–2264.