Identify Latent Group Structures in Panel Data: The classifylasso Command

Wenxin Huang[1] Yiru Wang[2] Lingyun Zhou[3]

[1]Shanghai Jiao Tong University

[2]University of Pittsburgh

[3]Tsinghua University

Canadian Stata conference, 2023

Motivation

$$y_{i,t} = \mu_i + \underbrace{\beta_i'}_{?} x_{i,t} + u_{i,t},$$

- Slope homogeneity?
- Slope heterogeneity?

Literature: slope heterogeneity

- Slope homogeneity: $\beta_i = \beta, \forall i = 1, \cdots, N.$ found to fail: Burnside (1996), Hsiao and Tahmiscioglu (1997), Lee, Pesaran, and Smith (1997), Durlauf, Kourtellos, and Minkin (2001), Phillips and Sul (2007a), Browning and Carro (2007), Browning and Carro (2010), Su and Chen (2013), Browning and Carro (2014), etc.
- Slope heterogeneity:
    - Complete slope heterogeneity.
    - Groups: the panel structure models individuals as belonging to a number of homogeneous groups or clubs within a broadly heterogeneous population.
        - Known group structure: Bester and Hansen (2013)
        - Unknown group structure:
          (i) Finite mixture models: Sun (2005), Kasahara and Shimotsu (2009), Browning and Carro (2011), Vogt and Linton (2020).
          (ii) K-means algorithm: Lin and Ng (2012), Sarafidis and Weber (2011), Bonhomme and Manresa (2015), Zhang, Wang and Zhu (2019).
          (iii) Classifier-Lasso: Su, Shi, and Phillips (2016), Huang, Jin and Su (2020), Mehrabani (2022).

Latent group structure

$$y_{i,t} = \mu_i + \underbrace{\beta_i'}_{?} x_{i,t} + u_{i,t},$$

- **Unobserved cross-sectional heterogeneity:**
  - In one increasingly popular framework, researchers use the latent group structure to characterize the cross-sectional heterogeneity, such that

$$\underbrace{\beta_i}_{\# N} = \underbrace{\sum_{k=1}^{K} \gamma_k 1\left\{i \in G_k\right\}}_{\# K}.$$

Software packages

- MATLAB package: Su, Shi, and Phillips (2016)
- R package: Gao and Shi (2021)
- Stata package: classifylasso

Methodology

Model: Unobserved Heterogeneity in β_i

- Consider the following panel model

$$y_{it} = \mu_i + x_{it}'\beta_i^0 + \epsilon_{it}$$

- We allow the true values of β_i, denoted as β_i^0, to follow a **latent group pattern**

$$\beta_i^0 = \sum_{k=1}^{K_0} \alpha_k^0 \mathbf{1}\left\{i \in G_k^0\right\} \tag{1}$$

where $\alpha_j^0 \neq \alpha_k^0$ for any $j \neq k$, $\cup_{k=1}^{K_0} G_k^0 = \{1, 2, \cdots, N\}$, and $G_j^0 \cap G_k^0 = \varnothing$ for any $j \neq k$. Let $N_k = \# G_k^0$ denote the cardinality of the set G_k^0.

- The unobserved parameter heterogeneity is a **joint problem**:
  - Model selection: what types of heterogeneity structures
  - Parameter estimation: how to obtain consistent and efficient estimator
- In the estimation procedure, we temporarily assume K_0 is known. In practice, we have IC to determine the $\#$ of groups.

The Classifier-Lasso Estimation

- The C-Lasso method **jointly** estimates **group-specific parameters** and identify the **unknown group membership.**
- **Intuition:** it shrinks the fully heterogeneous parameter β_i into the group-specific one α_k,

$$\{\widehat{\beta}_i, \widehat{\alpha}_k\} = \arg\min_{\beta_i, \alpha_k} \left(\underbrace{Q_{NT}(\beta)}_{\text{Loss}} + \underbrace{\frac{\lambda}{N} \sum_{i=1}^{N} \prod_{k=1}^{K} \|\beta_i - \alpha_k\|}_{\text{Penalty}} \right).$$

- The C-Lasso approach maintains the core insight of Lasso – *parameter sparsity*.
    - Limited heterogeneity: achieve efficiency within a group.
    - Data-driven model selection: unknown group patterns.

Estimation Procedure

1. **Initial estimation.** Obtain initial estimates of β_i from the OLS method.
2. **C-Lasso estimation.** Minimize the penalized-LS criterion function to obtain the C-Lasso estimates $\widehat{\beta}$ and $\widehat{\alpha}$ and the estimated groups $\widehat{G}_k = \left\{ i \in \{1, 2, \ldots, N\} : \widehat{\beta}_i = \widehat{\alpha}_k \right\}$ for $k = 1, \cdots, K$.
3. **Post-Lasso estimation.** Given the estimated group memberships, obtain the post-Lasso estimators $\widehat{\alpha}^{post}_{\widehat{G}_k}$.
4. **Group number selection.** Minimize the following BIC-type information criterion to select the number of groups $\widehat{K}$.

$$IC(\widehat{K}) = ln\left(\widehat{\sigma}^2_{\widehat{G}(\widehat{K})}\right) + \rho p \widehat{K},$$

where $\widehat{\sigma}^2_{\widehat{G}(\widehat{K})} = \frac{1}{NT} \sum_{k=1}^{\widehat{K}} \sum_{i \in \widehat{G}_k(\widehat{K})} \sum_{t=1}^{T} \left(\tilde{y}_{it} - \widehat{\alpha}^{post\prime}_{\widehat{G}_k(\widehat{K})} \tilde{x}_{it} \right)^2$.

classifylasso: Syntax

classifylasso - Identify latent group structures in panel data.

classifylasso *depvar indepvar* [*if*] [*in*] [, *options*]

options	
group(*numlist*)	specifies the possible number (list) of latent groups
lambda(#)	specifies the constant c_λ in $\lambda_{NT} = c_\lambda T^{-1/3}$
rho(#)	specifies the constant c_ρ in $\rho_{NT} = c_\rho (NT)^{-1/2}$
tolerance(#)	specifies the tolerance criterion for convergence
maxiteration(#)	specifies the maximum level of iterations
optimize_options	control the optimize package
absorb(*varlist*)	specifies the categorical variables of the fixed effects
noabsorb	suppresses the fixed effects.
vce(*vcetype*)	specifies the standard error type in post-Lasso estimation
dynamic	applies half-panel jackknife method to correct bias
notable	suppresses the estimation table
display_options	control the display style

Postestimation commands

- `classoselect`: determines the active result to be used in the following `predict`, `estimates replay` and `classocoef`;

  `classoselect` , *options*

options	
`group(#)`	specifies the number of groups use
`postselection`	specifies the post-Lasso estimation results
`penalized`	specifies the C-Lasso estimation results

- `predict`: generates new variables containing group membership, fitted values, and residuals;

  `predict` *newvar* [*if*] [*in*] [, *statistic*]

statistic	
`gid`	predicts the group membership, and it is the default
`xb`	predicts the linear prediction
`d`	calculates the fixed effects
`xbd`	predicts the sum of `xb` and `d`
`residuals`	calculates the residuals
`stdp`	calculates the standard deviation of linear prediction

Postestimation commands

- `estimates replay`: displays and exports the table of coefficient estimtes;
  `estimates replay` [, *options*]

options	
display_options	control the display style
`outreg2(`*filename* [, *options*]`)`	exports the coefficients to local disk

- `classocoef`: visualizes the coefficients in graphs;
  `classocoef` [*indepvar*] [, *options*]

- `classogroup`: plots the group number selection information.
  `classogroup` [, *options*]

Empirical Study

Empirical 1: determinants of savings

- Su, Shi, and Phillips (2016): the determinants of savings through a balanced panel of 56 countries from 1995 to 2010.

- **Regression model:**

$$\begin{aligned} Saving_{it} = \beta_{1i} Saving_{i,t-1} + \beta_{2i} \% \Delta CPI_{it} + \beta_{3i} Interest_{it} \\ + \beta_{4i} \% \Delta GDP_{it} + \mu_i + u_{it}, \end{aligned}$$

- Setting: tuning parameter $c_\lambda = 1.5485$, use the dynamic panel, and select the group numbers from 1 to 5.

Empirical 1: determinants of savings

```
use saving.dta, clear
xtset code year
classifylasso savings lagsavings cpi interest gdp, ///
    group(1/5) lambda(1.5485) tol(1e-4) dynamic

** Process of the iterative algorithm
Estimation 1: Group Number = 1; Iteration: ✓
Information Criterion = -.359883766
Estimation 2: Group Number = 2; Iteration: 1···5····10····15····20✓
Information Criterion = -.369981214
Estimation 3: Group Number = 3; Iteration: 1···5····10····15····20✓
Information Criterion = -.302279905
Estimation 4: Group Number = 4; Iteration: 1···5····10····15····20✓
Information Criterion = -.208381894
Estimation 5: Group Number = 5; Iteration: 1···5····10····15····20✓
Information Criterion = -.069495226
* Selected Group Number: 2
The algorithm takes 7min25s.
```

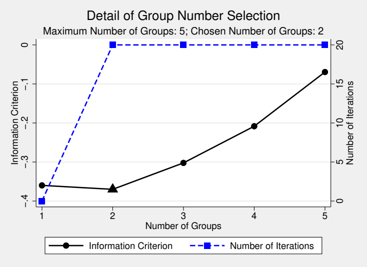
Empirical 1: determinants of savings

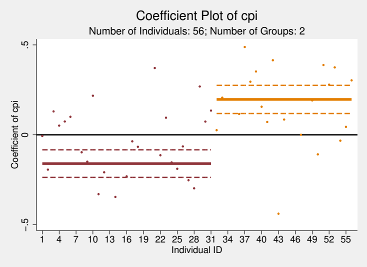
```
** Estimation table
Classifier-Lasso linear model                     Number of obs    =        840
Postestimation with 2 groups                      Number of units =          56

Fixed effect estimation with Group 1              R-squared        =     0.4988
Absorbing: code                                   Adj R-squared    =     0.4592
No. of obs   =    465                             Within R-sq.     =     0.4988
No. of units =     31                             Root MSE         =     0.7362
------------------------------------------------------------------------------
     savings | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
  lagsavings |   .6952103   .0383023    18.15   0.000     .6201392    .7702815
         cpi |   -.160168    .039182    -4.09   0.000    -.2369634   -.0833727
    interest |  -.1490145   .0368407    -4.04   0.000     -.221221    -.076808
         gdp |   .2892251   .0379408     7.62   0.000     .2148624    .3635878
       _cons |   .0550013   .0320203     1.72   0.086    -.0077574      .11776
------------------------------------------------------------------------------
Fixed effect estimation with Group 2              R-squared        =     0.4372
Absorbing: code                                   Adj R-squared    =     0.3917
No. of obs   =    375                             Within R-sq.     =     0.4372
No. of units =     25                             Root MSE         =     0.7810
------------------------------------------------------------------------------
     savings | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
  lagsavings |   .6938863   .0356796    19.45   0.000     .6239556     .763817
         cpi |   .1967192   .0399412     4.93   0.000     .1184359    .2750025
    interest |   .1225496   .0411717     2.98   0.003     .0418545    .2032447
         gdp |   .1126528   .0474176     2.38   0.018     .0197161    .2055896
       _cons |  -.0067423   .0401331    -0.17   0.867    -.0854018    .0719172
------------------------------------------------------------------------------
```

Empirical 1: determinants of savings

```
** Visualization
classogroup, export("selection1.eps") // group selection information
classocoef cpi, export("coefcpi.eps") // coefficient plot
```





Empirical 2: democracy and economic growth

- Acemoglu et al. (2019): the relationship between democracy and economic growth
- **Regression model:**

$$lnPGDP_{it} = \beta_i Democracy_{it} + \sum_{j=1}^{l} \gamma_{i,j} lnPGDP_{i,t-j} + \mu_i + \lambda_t + u_{it},$$

- To obtain robust results, we consider the specifications including 1, 2, 3, 4 lags, respectively.

Empirical 2: democracy and economic growth

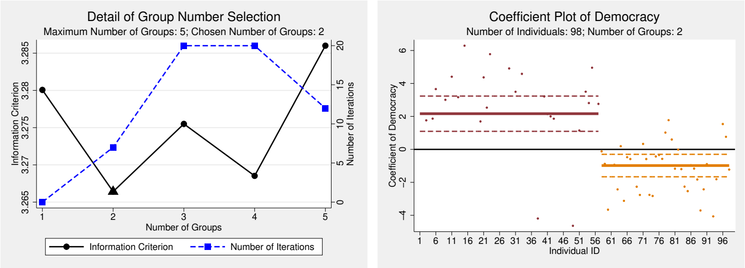


Figure: Heterogeneous Effects of Democracy on Economic Growth

Empirical 2: democracy and economic growth

Table: Heterogeneous Effects of Democracy on Economic Growth

		(1)			(2)			(3)			(4)	
$lnPGDP$	Pooled	G1	G2	Pooled	G1	G2	Pooled	G1	G2	Pooled	G1	G2
Democracy	1.055	2.165	-0.981	0.781	1.622	-0.869	0.763	1.089	-1.462	0.842	1.165	-1.172
	(0.370)	(0.545)	(0.348)	(0.263)	(0.339)	(0.365)	(0.259)	(0.314)	(0.305)	(0.258)	(0.313)	(0.303)
$lnPGDP_{-1}$	0.970	1.033	0.982	1.250	1.309	1.333	1.227	1.335	1.133	1.228	1.347	1.088
	(0.006)	(0.007)	(0.009)	(0.062)	(0.075)	(0.126)	(0.055)	(0.066)	(0.057)	(0.057)	(0.068)	(0.056)
$lnPGDP_{-2}$				-0.284	-0.287	-0.314	-0.194	-0.223	-0.142	-0.214	-0.250	-0.131
				(0.061)	(0.074)	(0.122)	(0.051)	(0.063)	(0.073)	(0.052)	(0.065)	(0.072)
$lnPGDP_{-3}$							-0.069	-0.072	-0.006	-0.006	-0.033	0.082
							(0.027)	(0.029)	(0.038)	(0.037)	(0.042)	(0.058)
$lnPGDP_{-4}$										-0.046	-0.027	-0.042
										(0.021)	(0.025)	(0.050)
Country FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
N	98	57	41	98	59	39	98	61	37	98	67	31
T	40	40	40	39	39	39	38	38	38	37	37	37
$\#Obs.$	3,920	2,280	1,640	3,822	2,301	1,521	3,724	2,318	1,406	3,626	2,479	1,147

Empirical 2: democracy and economic growth

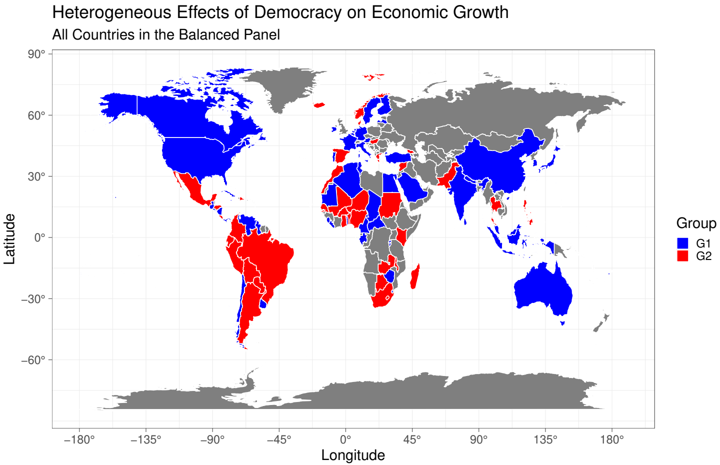

Heterogeneous Effects of Democracy on Economic Growth
All Countries in the Balanced Panel
90°
60°
30°
0°
−30°
−60°
Latitude
−180°
−135°
−90°
−45°
0°
45°
90°
135°
180°
Longitude
Group
🟦 G1
🟥 G2

Figure: Heterogeneous Effects of Democracy in the World

Simulations

DGP

- linear static panels with latent group structures
- three groups with proportion $N_1 : N_2 : N_3 = 0.3 : 0.3 : 0.4$
- sample size: $N \in \{100, 200\}$, $T \in \{20, 40\}$, $p \in \{2, 4\}$
- true coefficients:
  - For $p = 2$:
    - $\alpha_1 = (0.4, 1.6)$
    - $\alpha_2 = (1, 1)$
    - $\alpha_3 = (1.6, 0.4)$
  - For $p = 4$:
    - $\alpha_1 = (0.4, 1.6, -0.4, -1.6)$
    - $\alpha_2 = (1, 1, -1, -1)$
    - $\alpha_3 = (1.6, 0.4, -1.6, -0.4)$
- tunings (use the default values in the command): $c_\lambda = 0.5$ for $\lambda_{NT} = c_\lambda T^{-1/3}$ and $c_\rho = 2/3$ for $\rho_{NT} = c_\rho (NT)^{-1/2}$
- we select the group number from 1 to 5

Simulation results: select K

Table: Selecting the number of groups

			Frequency of selecting K					Computation time (minutes)				
N	T	p	1	2	**3**	4	5	1	2	3	4	5
100	20	2	0	0	**0.998**	0.002	0	0.004	0.391	0.648	0.821	1.015
100	40	2	0	0	**1**	0	0	0.010	0.601	0.900	1.052	0.735
100	20	4	0	0	**0.99**	0.01	0	0.012	1.120	1.907	2.388	2.878
100	40	4	0	0	**1**	0	0	0.039	1.582	2.261	3.009	2.145
200	20	2	0	0	**0.998**	0.002	0	0.004	0.377	0.949	2.163	2.925
200	40	2	0	0	**1**	0	0	0.008	0.432	1.117	2.662	2.182
200	20	4	0	0	**1**	0	0	0.012	1.544	4.157	8.837	12.826
200	40	4	0	0	**1**	0	0	0.039	1.858	4.830	12.267	14.653

Simulation results: estimation accuracy

Table: Classification accuracy and estimation performance of α_1

			Correct	Post-Lasso			Oracle		
N	T	p	Classification	RMSE	Bias	Coverage	RMSE	Bias	Coverage
100	20	2	0.9354	0.0446	0.0114	0.9068	0.0383	-0.0014	0.9538
100	40	2	0.9900	0.0274	0.0024	0.9442	0.0266	0.0004	0.9488
100	20	4	0.9392	0.0321	0.0124	0.8942	0.0266	0.0006	0.9548
100	40	4	0.9899	0.0195	0.0013	0.9398	0.0189	-0.0006	0.9476
200	20	2	0.9785	0.0417	0.0058	0.9326	0.0391	0.0007	0.9494
200	40	2	0.9990	0.0275	0.0001	0.9362	0.0274	-0.0002	0.9370
200	20	4	0.9775	0.0298	0.0047	0.9254	0.0276	-0.0003	0.9436
200	40	4	0.9992	0.0193	-0.0001	0.9490	0.0192	-0.0003	0.9484

Simulation results: beta-min assumption

We consider DGP of two covariates ($p = 2$) and groupwise parameters

- $\alpha_1 = (1 - C, 1 + C)$
- $\alpha_2 = (1, 1)$
- $\alpha_3 = (1 + C, 1 - C)$

with values of $C \in \{0.01, 0.1, 0.3, 0.6\}$.

Table: Classification and performance under violation of the beta-min assumption

		Correct Classification				Coverage of α_1			
N	T	$C = 0.01$	0.1	0.3	0.6	0.01	0.1	0.3	0.6
100	20	0.3997	0.4719	0.7428	0.9354	0.2616	0.3428	0.8026	0.9068
100	40	0.4019	0.5246	0.8593	0.9900	0.3412	0.4886	0.8644	0.9442
200	20	0.3808	0.4720	0.7473	0.9392	0.1494	0.2354	0.7596	0.8942
200	40	0.3815	0.5038	0.8597	0.9899	0.1880	0.3048	0.8146	0.9398

Thanks for your attention!