Documentation of the Panel Data Models with Fixed Effects project

Yuxin Wang

29 March 2021

CONTENTS

CHAPTER ONE

INTRODUCTION

In this project we present the main ideas of Bai (2009), Bai and Ng (2002), and Moon and Weidner (2015). We start with comparing the convergence properties and identification conditions for least squares estimators in panel data models, and then use econometric theory for factor model estimations in practice. We provide the algorithms in Bai (2009) and Bai, Ng (2002) in Python, replicate some Monte Carlo simulation results of the papers and investigate the simulation with other data generating processes and visualize them in R Language. In the end, we return to the previous empirical example and look at the estimation using the estimator proposed by Bai (2009).

1.1 Project structure

The logic of this project works by step of the analysis:

1. Data management
2. Monte Carlo simulations and real data application
3. Visualisation and results formatting
4. Research paper and presentation

The main part of this project is Monte Carlo simulations and real data estimation. See section *Monte Carlo Simulations and Real Data Application* for more detail.

1.2 Getting started

This assumes you have installed:

- Miniconda or Anaconda
- a modern LaTeX distribution (e.g. TeXLive, MacTex, or MikTex)

1. Create the environment from the *environment.yml* file.

```
$ conda env create -f environment.yml
```

2. Activate the newly created conda environment.

```
$ conda activate panel_data_models_with_fixed_effects
```

3. Run pytask to start all tasks automaticly.

```
$ conda develop .
$ pytask
```

After that, all build result can be found in *bld* folder.

1.3 Built result

The main analysis task takes hours to run. For quickly scan, some built result can be found in this repository: panel_data_models_build_result

CHAPTER

TWO

ORIGINAL DATA

Documentation of the different datasets in *original_data*.

2.1 Cigar

Baltagi and Levin (1992) studied the demand for cigarettes in US states using a panel data model.

The data set is a panel of 46 observations from 1963 to 1992.

Containing :

- state: state abbreviation
- year: the year
- price: price per pack of cigarettes
- pop: population
- pop16: population above the age of 16
- cpi: consumer price index (1983=100)
- ndi: per capita disposable income
- sales: cigarette sales in packs per capita
- pimin: minimum price in adjoining states per pack of cigarettes

CHAPTER

THREE

DATA MANAGEMENT

Documentation of the code in *src/data_management*.

3.1 Cigar Data Processing

Compute the logarithm of all variables and adjust the monetary variables for the general consumer price index. Also compute a lagged consumption variable.

CHAPTER

FOUR

MONTE CARLO SIMULATIONS AND REAL DATA APPLICATION

Documentation of the code in *src.analysis*. This is the core of the project.

The analysis in this project contains three aspects:

1. **Estimation of slope coefficients**
    1. Monte carlo simulations under IID assumptions
    2. Monte carlo simulations under ARI time fixed effects
    3. Starting Values in the additive fixed effects model
2. **Estimation of the number of factors**
    1. Estimate parameters in interactive fixed effects model by using different numbers of factors
    2. Estimate factor numbers in interactive fixed effects model by choosing different penalty functions and criterias
3. **Real data application**
    1. Use a panel data set of 46 observations from 1963 to 1992
    2. Estimate the coefficients and factor numbers by using the model in Baltagi and Levin (1992)

Process of Monte Carlo simulations:

1. Data generating processes for monte carlo simulations
2. Estimate model parameters by interactive-effects estimator and within estimator
3. Caculate statistical results for coefficients, including mean, bias, root-mean-square error, standard error and confidence interval

Function *simulation_coefficient* in module *src.analysis.simulation* is defined for running the simulations under different number of individuals and time periods over 1000 repetitions.

4.1 Data Generating process for Monte Carlo Simulation

dgp_additive_fixed_effects_model(*T*, *N*, *, *beta1*, *beta2*, **kw*)

Data generating process for "Interactive Fixed Effects Model with Common Regressors and Time-invariant Regressors"

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

X [array-like] Simulate data of exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \alpha_i + \xi_t + \epsilon_{it}$$

Two fixed effects satisfy

$$\alpha_i, \xi_t \overset{\text{i.i.d}}{\sim} N(0,1).$$

Both of two fixed effects are correlated with the two regressors

$$X_{it,j} = 3 + 2\alpha_i + 2\xi_t + \eta_{it,j},$$

with

$$\eta_{it,j} \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

The error term satisfies

$$\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0,4).$$

dgp_additive_fixed_effects_model_no_iid(*T*, *N*, *, *beta1*, *beta2*, **kw*)

Monte carlo data generate processor for "Interactive Fixed Effects Model with Common Regressors and Time-invariant Regressors"

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

X [array-like] Simulate data of Exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of Dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \alpha_i + \xi_t + \epsilon_{it}$$

Two fixed effects satisfy

$$\alpha_i \overset{\text{i.i.d}}{\sim} N(0,1).\xi_t = 0.7 \cdot \xi_{t-1} + u_t,$$

with:

$$u_t \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

Both of two fixed effects are correlated with the two regressors

$$X_{it,j} = 3 + 2\alpha_i + 2\xi_t + \eta_{it,j},$$

with

$$\eta_{it,j} \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

The error term satisfies

$$\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0,4).$$

dgp_interactive_fixed_effects_model(*T*, *N*, *, *beta1*, *beta2*, *mu*, **kw*)

Data generating process for "Interactive Fixed Effects Model"

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

mu [float] Constant

X [array-like] Simulate data of exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \mu + \lambda_i' F_t + \epsilon_{it}$$

where

$$\lambda_i = \begin{pmatrix}\lambda_{i1}\\ \lambda_{i2}\end{pmatrix} \overset{\text{i.i.d}}{\sim} N(0,I_2), F_t = \begin{pmatrix}F_{t1}\\ F_{t2}\end{pmatrix} \overset{\text{i.i.d}}{\sim} N(0,I_2),$$

The regressors are generated according to:

$$X_{it,j} = 1 + \lambda_{i1}F_{t1} + \lambda_{i2}F_{t2} + \lambda_{i1} + \lambda_{i2} + F_{t1} + F_{t2} + \eta_{it,j},$$

with

$$\eta_{it,j} \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

The regressors are correlated with λ_i, f_t, and the product $\lambda_i' F_t$. The error term satisfies

$$\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0,4).$$

dgp_interactive_fixed_effects_model_no_iid (*T*, *N*, *, *beta1*, *beta2*, *mu*, ***kw*)
Monte carlo data generate processor for "Interactive Fixed Effects Model"

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

mu [float] Constant

X [array-like] Simulate data of Exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of Dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \mu + \lambda_i' F_t + \epsilon_{it}$$

where

$$\lambda_i = \begin{pmatrix} \lambda_{i1} \\ \lambda_{i2} \end{pmatrix} \overset{\text{i.i.d}}{\sim} N(0, I_2), F_{t,j} = 0.7 \cdot F_{t-1,j} + u_{t,j},$$

with:

$$u_{t,j} \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

The regressors are generated according to:

$$X_{it,j} = 1 + \lambda_{i1} F_{t1} + \lambda_{i2} F_{t2} + \lambda_{i1} + \lambda_{i2} + F_{t1} + F_{t2} + \eta_{it,j},$$

with

$$\eta_{it,j} \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

The regressors are correlated with λ_i, f_t, and the product $\lambda_i' F_t$. The error term satisfies

$$\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0,4).$$

dgp_interactive_fixed_effects_model_with_common_and_time_invariant (*T*, *N*, *, *beta1*, *beta2*, *mu*, *gamma*, *delta*, ***kw*)
Data generating process for "Interactive Fixed Effects Model with Common Regressors and Time-invariant Regressors"

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

mu [float] Constant

gamma [float] Coefficient of x_i

delta [float] Coefficient of w_t

X [array-like] Simulate data of exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \mu + x_i\gamma + w_t\delta + \lambda_i' F_t + \epsilon_{it}$$

where

$$\lambda_i = \begin{pmatrix}\lambda_{i1}\\ \lambda_{i2}\end{pmatrix} \overset{\text{i.i.d}}{\sim} N(0, I_2), F_t = \begin{pmatrix}F_{t1}\\ F_{t2}\end{pmatrix} \overset{\text{i.i.d}}{\sim} N(0, I_2).$$

The regressors are generated according to:

$$X_{it,j} = 1 + \lambda_{i1}F_{t1} + \lambda_{i2}F_{t2} + \lambda_{i1} + \lambda_{i2} + F_{t1} + F_{t2} + \eta_{it,j},$$

with

$$\eta_{it,j} \overset{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

Additionally, we set

$$x_i = \lambda_{i1} + \lambda_{i2} + e_i, \qquad e_i \overset{\text{i.i.d}}{\sim} N(0,1)$$

and

$$w_t = F_{t1} + F_{t2} + \eta_t, \qquad \eta_t \overset{\text{i.i.d}}{\sim} N(0,1),$$

so that x_i is correlated with λ_i and w_t is correlated with f_t.

dgp_interactive_fixed_effects_model_with_common_and_time_invariant_no_iid(*T*, *N*, ***, *beta1*, *beta2*, *mu*, *gamma*, *delta*, ***kw*)

Monte carlo data generate processor for “Interactive Fixed Effects Model with Common Regressors and Time-invariant Regressors”

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

mu [float] Constant

gamma [float] Coefficient of x_i

delta [float] Coefficient of w_t

X [array-like] Simulate data of Exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of Dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \mu + x_i\gamma + w_t\delta + \lambda_i' F_t + \epsilon_{it}$$

where

$$\lambda_i = \begin{pmatrix}\lambda_{i1}\\ \lambda_{i2}\end{pmatrix} \stackrel{\text{i.i.d}}{\sim} N(0, I_2), F_{t,j} = 0.7 \cdot F_{t-1,j} + u_{t,j},$$

with:

$$u_{t,j} \stackrel{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

The regressors are generated according to:

$$X_{it,j} = 1 + \lambda_{i1}F_{t1} + \lambda_{i2}F_{t2} + \lambda_{i1} + \lambda_{i2} + F_{t1} + F_{t2} + \eta_{it,j},$$

with

$$\eta_{it,j} \stackrel{\text{i.i.d}}{\sim} N(0,1) \qquad j \in \{1,2\}.$$

Additionally, we set

$$x_i = \lambda_{i1} + \lambda_{i2} + e_i, \qquad e_i \stackrel{\text{i.i.d}}{\sim} N(0,1)$$

and

$$w_t = F_{t1} + F_{t2} + \eta_t, \qquad \eta_t \stackrel{\text{i.i.d}}{\sim} N(0,1),$$

so that x_i is correlated with λ_i and w_t is correlated with f_t.

`dgp_random_iid_residual`(*N*, *T*, *r*)
Data generating process for random distributed residual.Residual = Y - beta * X = Lambda * Factor + epsShape of residual is (N, T).

`dgp_time_invariant_fixed_effects_model`(*T*, *N*, *, *beta1*, *beta2*, ***kw*)
Data generating process for "Interactive Fixed Effects Model with Common Regressors and Time-invariant Regressors"

T [int] Sample size of time

N [int] Sample size of entity

beta1 [float] Coefficient of x1

beta2 [float] Coefficient of x2

X [array-like] Simulate data of exogenous or right-hand-side variables (variable by time by entity).

Y [array-like] Simulate data of dependent (left-hand-side) variable (time by entity).

$$y_{it} = \beta_1 x_{it,1} + \beta_2 x_{it,2} + \alpha_i + \epsilon_{it}$$

The regressors are generated according to $X_{it,j} = 3 + 2\alpha_i + \eta_{it,j}$, with

$$\eta_{it,j} \overset{\text{i.i.d}}{\sim} N(0,1), \qquad j \in \{1,2\},$$
$$\alpha_i \overset{\text{i.i.d}}{\sim} N(0,1),$$
$$\epsilon_{it} \overset{\text{i.i.d}}{\sim} N(0,4).$$

4.2 Monte Carlo Simulation

`simulation_coefficient`(*dgp_func*, *all_N*, *all_T*, *nsims*, ***, *need_sde=False*, *tolerance=0.0001*, *r=2*, *interactive_start_value_effect='pooling'*, *within_effect='twoways'*, ***beta_true*)

Monte carlo simulation to estimate beta hat and standard error of coefficient.

dgp_func [function] One function in monte_carlo_dgp

all_N [array-like] Different sample sizes of entity

all_T [array-like] Different sample sizes of time

nsims [int] Simulation times under the same N and T

need_sde [bool] Flag the sde caculation conditions

tolerance [float, optional] Iteration precision.

r [int, optional] Number of factors.

interactive_start_value_effect [string, optional] The effects used in package linearmodels for starting value, one of "pooling", "twoways"

within_effect [string, optional] The effects used in package linearmodels for within estimator, one of "twoways", "individual"

beta_true [float] Coefficient of variables used in dgp_func. Values in ("beta1", "beta2", "mu", "gamma", "delta")

df_sim_result [DataFrame] Columns are T, N, sim, beta_interactive, beta_within, sde_interactive, sde_within

`statistics_coefficient`(*all_N*, *all_T*, *nsims*, *df_sim_result*, ***beta_true*)

Generate statistics of each N & T, take the mean of different simulations, and store them in a data frame. We include mean, bias, the RMSE, standard error and cofidence interval in our statistical results.

all_N [array-like] Different sample sizes of entity

all_T [array-like] Different sample sizes of time

nsims [int] Simulation times under the same N and T

df_sim_result [DataFrame] Simulation results from function *simulation_coefficient*

beta_true [float] Coefficients of variables used in dgp_func. Values in ("beta1", "beta2", "mu", "gamma", "delta")

4.3 Estimation of slope coefficients

Task for applying different models, number of individuals and time periods to function *simulation_coefficient*.

4.4 Statistics

Task for caculating statistical results for coefficients by using *statistics_coefficient* function. The inputs of the function come from *task_simulation_coefficient*.

4.5 Estimation of the number of factors

Task for estimating the parameters in interactive fixed effects model by using different numbers of factors.Task for estimating factor numbers in interactive fixed effects model by choosing different penalty functions g1,g2,g3 with criterias PC and IC.

`task_factor_estimate_interactive_fixed_effects_model` (*produces*)
: Task for estimating factor numbers in interactive fixed effects model. We choose different penalty functions g1,g2,g3 with criterias PC and IC.

`task_factor_estimate_random_iid_residual` (*produces*)
: Task for estimating factor numbers in the model defined by the function *dgp_random_iid_residual*. We choose different penalty functions g1,g2,g3 with criterias PC and IC. It replicates the result of Table 2 in Bai,Ng (2002), page 205.

`task_simulation_and_statistics_range_r` (*depends_on*, *produces*)
: Task for estimating the parameters in interactive fixed effects model by using different numbers of factors.

CHAPTER

FIVE

VISUALISATION AND RESULTS FORMATTING

Documentation of the code in *src.final*.

Execute R-script by pytask-r, for generating tables and figures used in our final paper.

Moreover, we also show that factor estimation models have good finite sample properties in many configurations of panel data sets.

5.1 Generating Figures for Coefficients

Use figures to visualize the convergence properties of the estimators mentioned in the paper, and compare how well they work in different panel data models.

5.2 Generating Tables for Coefficients

Use tables to show the convergence properties of the estimators mentioned in the paper, and compare how well they work in different panel data models.

5.3 Generating Figures and Tables for Factor Estimation

Use figures and tables to show the properties of factor estimation models in finite panel data sets.

CHAPTER

SIX

MODEL CODE

This directory *src.model_code* contains interactive fixed effects model estimator and factor number estimator, and some statistical and utility funcitons.

6.1 The `InteractiveFixedEffect` class

Calculate the interactive fixed effect estimator as described in Bai(2009). The corresponding theory is in Chapter 3 of our report.

`class InteractiveFixedEffect`(*dependent*, *exog*)

Interactive fixed effects estimator for panel data

dependent [array-like] Dependent (left-hand-side) variable (time by entity).

exog [array-like] Exogenous or right-hand-side variables (variable by time by entity).

$$y_{it} = \beta x_{it} + \lambda_i' F_t + \epsilon_{it}$$

`calculate_sde`(*beta_hat*, *f_hat*, *lambda_hat*)

Calculate standard error of beta_hat estimated from fit

`fit`(*r*, *beta_hat_0=None*, *tolerance=0.0001*)

Estimate parameters of different model

r [int] Number of factors.

beta_hat_0: array-like, optional Starting values of estimator. Should be same order as exog variable.

tolerance [float, optional] Iteration precision.

beta_hat [array-like] Estimate the result of slope coefficients. Same order as exog variable.

beta_hat_list [array-like] Iteration intermediate values.

f_hat [array-like] Estimate the result of time fixed effects.

lambda_hat [array-like] Estimate the result of individual fixed effects.

6.2 The `Statistics` module

Some pure functions to caculate statistical results

`caculate_rmse` (*estimated*, *real*)
Caculate the root-mean-square-error

estimated [array-like] Estimated value

real [array-like] Real value

`caculate_sde` (*estimated*)
Caculate standard error

estimated [array-like] Estimated value

6.3 The `utils` module

Some pure functions used frequently.

`paste` (**lists*, *sep=' '*, *collapse=None*)
Concatenate vectors after converting to character. Implement the function *Paste* in R.

lists [array-like] One or more objects, to be converted into character vectors.

sep [string] A character string used to separate the terms

collapse [string] An optional character string used to separate the results

`paste0` (**lists*, *collapse=None*)
paste0(…, collapse) is similar to paste(…, sep = "", collapse), but the former one is slightly more efficient.

`reduce_concat` (*x*, *sep=''*)
Join a list into one string.

`stretch` (*lists*)
Stretch lists to same length. Shorter lists is repeated to reach the same length as the longest.

CHAPTER

SEVEN

MODEL SPECIFICATIONS

The directory *src.model_specs* contains JSON files with model specifications.

These json files contain parameters for *Monte Carlo Simulation*. They are used in *Estimation of slope coefficients* to determine which panel data model to be used in each simulation.

CHAPTER EIGHT

RESEARCH PAPER / PRESENTATIONS

Purpose of the different files:

- `paper.tex` contains the actual paper.
- `slides.tex` contains a typical conference presentation.

CHAPTER

NINE

REFERENCES

- Bai, J. (2009). Panel data models with interactive fixed effects. Econometrica, 77(4), 1229-1279.
- Bai, J., & Ng, S. (2002). Determining the number of factors in approximate factor models. Econometrica, 70(1), 191-221.
- Baltagi, B. H., & Levin, D. (1992). Cigarette taxation: Raising revenues and reducing consumption. Structural Change and Economic Dynamics, 3(2), 321-335.
- Greene, W. H. (2003). Econometric analysis. Pearson Education.
- Holtz-Eakin, D., Newey, W., & Rosen, H. S. (1988). Estimating vector autoregressions with panel data. Econometrica: Journal of the econometric society, 1371-1395.
- Moon, H. R., & Weidner, M. (2015). Linear regression for panel with unknown number of factors as interactive fixed effects. Econometrica, 83(4), 1543-1579.
- Pesaran, M. H. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure. Econometrica, 74(4), 967-1012.
- Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. Journal of the American statistical association, 97(460), 1167-1179.
- Su, L., & Jin, S. (2012). Sieve estimation of panel data models with cross section dependence. Journal of Econometrics, 169(1), 34-47.
- Westerlund, J. (2019). On estimation and inference in heterogeneous panel regressions with interactive effects. Journal of Time Series Analysis, 40(5), 852-857.

PYTHON MODULE INDEX

a

d

f

m

INDEX

T