Linear Models with Multi–Way Fixed Effects

Sergio Correia, Federal Reserve Board

March 1, 2017

sergio.a.correia@frb.gov

Agenda

- Rationale for MWFE
- Implementation: `reghdfe`
- Estimator (Correia 2017)
    - Benchmarks
- Warnings:
    - Efficiency
    - FEs as objects of interest (Arora, Belenzon, Correia WiP 2017)
    - Inference with robust SEs (Cattaneo et al 2017, Verdier 2017)

What are Linear MWFE Models

We want to compute the least squares estimates $\hat{\boldsymbol{\beta}}$ of

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

- $\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 & \mathbf{D}_2 & \cdots & \mathbf{D}_F \end{bmatrix}$ consists of F indicator matrices
- If $F = 1$, this collapses to a standard fixed effect regression (`xtreg`, `areg`)
- Can't use dummies because $\begin{bmatrix} \mathbf{D}_2 & \cdots & \mathbf{D}_F \end{bmatrix}$ is too large

Why use Linear MWFE Models

- Control for unobservables that stay constant within an economic unit (workers, firms, exporters, importers, etc.)
- Example:
    - Want to study role of tax incentives on dividend payout policy
    - Payout policy differs vastly by industry, is affected by CEO preferences, and by current economic conditions.
    - How to control for that? add industry, CEO and time FEs.
- Applications in many fields: accounting (DeHaan et al 2015), finance (Gormley et al 2015), labor (Guimarães et al 2015), trade (Mayer 2016), etc.
- Alternatives: pooled cross–sections (possibly biased/inconsistent); random and mixed models (same).

Stata Implementation: `reghdfe`

```
reghdfe dividends tax_benefits,
     absorb(ceo industry year) vce(cluster ceo firm)

reghdfe dividends tax_benefits,
     absorb(industry#year ceo)

reghdfe dividends i.firm_type##c.tax_benefits,
     absorb(industry#year ceo)
```

Stata Implementation: `reghdfe`

- Allows other linear models: IV, GMM
- It's a building block for interactive fixed effects models (Bai 2009), Poisson FEs (Guimaraes & Portugal 2015), structural gravity models (Zylkin 2016), etc.
- Supports multi-way clustering (Cameron et al) and many types of standard errors (Newey-West, HAC, Kiefer, etc.)
- Degrees-of-freedom corrections (count collinear FEs, drop singleton groups)
- Very fast (uses Mata and the `ftools` package)

Stata Implementation: reghdfe

```
. reghdfe dividends tax_benefits, absorb(ceo industry year) vce(cluster ceo firm)
(converged in 6 iterations)

HDFE Linear regression                            Number of obs   =      1,000
Absorbing 3 HDFE groups                           F(   1,     39) =       0.89
Statistics robust to heteroskedasticity           Prob > F        =     0.3511
                                                  R-squared       =     0.0715
                                                  Adj R-squared   =     0.0079
Number of clusters (ceo)     =         50         Within R-sq.    =     0.0007
Number of clusters (firm)    =         40         Root MSE        =     0.9645

                                (Std. Err. adjusted for 40 clusters in ceo firm)
------------------------------------------------------------------------------
             |               Robust
   dividends |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
tax_benefits |   .0241443   .0255819     0.94   0.351    -.0276001    .0758886
------------------------------------------------------------------------------

Absorbed degrees of freedom:
-----------------------------------------------------+
 Absorbed FE | Categories  - Redundant  = Num. Coefs |
-------------+---------------------------------------|
         ceo |         50          50           0   *|
    industry |         10           0          10    |
        year |          5           1           4    |
-----------------------------------------------------+
* = FE nested within cluster; treated as redundant for DoF computation
```

Figure 1: reghdfe screenshot

Econometric Estimator

Steps:

1. Compute the residuals of $\mathbf{y}$ and $\mathbf{X}$ against $\mathbf{D}$:

$$\tilde{\mathbf{y}} = \mathbf{M}_{\mathbf{D}}\mathbf{y}$$

$$\tilde{\mathbf{X}} = \mathbf{M}_{\mathbf{D}}\mathbf{X}$$

2. Apply the Frisch–Waugh–Lovell Theorem:

$$\hat{\boldsymbol{\beta}} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$$

Thus, we can just focus on one variable at a time: $\tilde{\mathbf{y}}$

(Note: $\mathbf{M}_{\mathbf{D}} = \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}$ is the residual–maker matrix)

Econometric Estimator

To obtain $\hat{\mathbf{y}} = \mathbf{M_D y}$, find an $\hat{\boldsymbol{\alpha}}$ that satisfies the normal equations

$$\mathbf{D}'\mathbf{e} = 0 \quad , \quad \mathbf{e} \overset{\text{def}}{=} \mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}}$$

In plain English:

> *For every level g of every fixed effect f the mean of the residuals must be zero:*

$$\overline{e_i} = 0 \quad , \quad i \in \mathcal{I}(f, g)$$

Note: We don't care if $\hat{\boldsymbol{\alpha}}$ is unique

Econometric Estimator

We can compute $\hat{\mathbf{y}}$ by two approaches:

1. Method of Alternating Projections (MAP), accelerated
2. A sparse linear equation solver (conjugate gradient or LSMR)

MAP - Definition

$$\lim_{n\to\infty} ||(\mathbf{M}_1 \cdot \mathbf{M}_2 \dots \mathbf{M}_F)^n \mathbf{y} - \mathbf{M}_{12\dots F}\mathbf{y}|| = 0$$

Suggests iteration:

$$\mathbf{y}_{k+1} = \underbrace{(\mathbf{M}_1 \cdot \mathbf{M}_2 \dots \mathbf{M}_F)}_{\text{Linear Transform } \mathbf{T}} \mathbf{y}_k$$

MAP - Approach

- We want to regress $\mathbf{y}$ against CEO and firm FEs, and compute the residuals
- Very easy to do for only one FE; just demean
- For two FEs:

```
areg y, absorb(ceo)
predict y, resid

areg y, absorb(firm)
predict y, resid

// Repeat until -y- converges
```

MAP - Problem #1

Bauschke et al (2003):

> *[...] The main practical drawback of the MAP appears to be that it is often slowly convergent [...] Franchetti and Light and Bauschke, Borwein, and Lewis have given examples showing that the convergence [...] can be arbitrarily slow!*

It can be very, very slow!

(In particular when the underlying fixed effects are *poorly connected*)

MAP - Solution #1

Guimarães & Portugal (2010) and Gaure (2013) apply accelerations that are related to steepest descent

$$\mathbf{y}_{k+1} = t \underbrace{(\mathbf{M}_1 \cdot \mathbf{M}_2 \ldots \mathbf{M}_F)}_{\text{Linear Transform } \mathbf{T}} \mathbf{y}_k + (1-t)\mathbf{y}_k$$

(In essence, this just extrapolates $\mathbf{y}_k$)

Often improve speeds significantly, but ...

MAP - Problem #2

Bauschke et al (2003):

> *[...] perhaps surprisingly, we show that the acceleration scheme may actually be slower than the MAP [...]!*

Hernández-Ramos et al (2011):

> *[...] the steepest descent method is known for its slowness in the presence of ill-conditioned problems [...]*

MAP - Solution #2

- Use a better acceleration: conjugate gradient (CG)
- But we need a twist: $T \stackrel{\text{def}}{=} \mathbf{M}_1 \cdot \mathbf{M}_2 \ldots \mathbf{M}_F$ is not symmetric and CG requires a symmetric transformation
- Solution: follow Hernández-Ramos et al (2011) and make it symmetric:

$$T^{\text{Sym}} \stackrel{\text{def}}{=} \mathbf{M}_1 \cdot \mathbf{M}_2 \ldots \mathbf{M}_F \ldots \mathbf{M}_2 \cdot \mathbf{M}_1$$

$$T^{\text{Cim}} \stackrel{\text{def}}{=} (\mathbf{M}_1 \cdot \mathbf{M}_2 \ldots \mathbf{M}_F)/F$$

- Theoretical advantages (monotonic convergence) and practical ones (almost as fast as other methods for easy problems, significantly faster for ill-defined ones)

Sparse Solver

- We can also apply a sparse solver directly, without computing normal equations (e.g. Abowd et al 2002, Gomez 2016)
- Equivalent to accelerated MAP but with a different scalar t
- Performance is (unsurprisingly) also similar to accelerated MAP (even if we precondition the sparse solver preconditioning)

Problem #3

- Both converge very slowly for datasets that are poorly connected

	y	id1	id2
1	1	0	0
2	0	0	1
3	0	1	1
4	0	1	2
5	0	2	2
6	0	2	3
7	0	3	3
8	0	3	4
9	0	4	4
10	0	4	5

Figure 2: This dataset will turn your PC into a heater in the winter

Solution #3: Link with Graph Theory

- Let's rewrite the two–way fixed effect model as a graph:

Indiv.	Firm	y
1	4	0.49
1	5	−1.41
2	4	−0.20
2	6	2.11
2	6	0.45
2	7	−0.32
3	7	0.76

Figure 3: Dataset for CEO–Firm regression

Solution #3: Link with Graph Theory

- Let's rewrite the two–way fixed effect model as a graph:

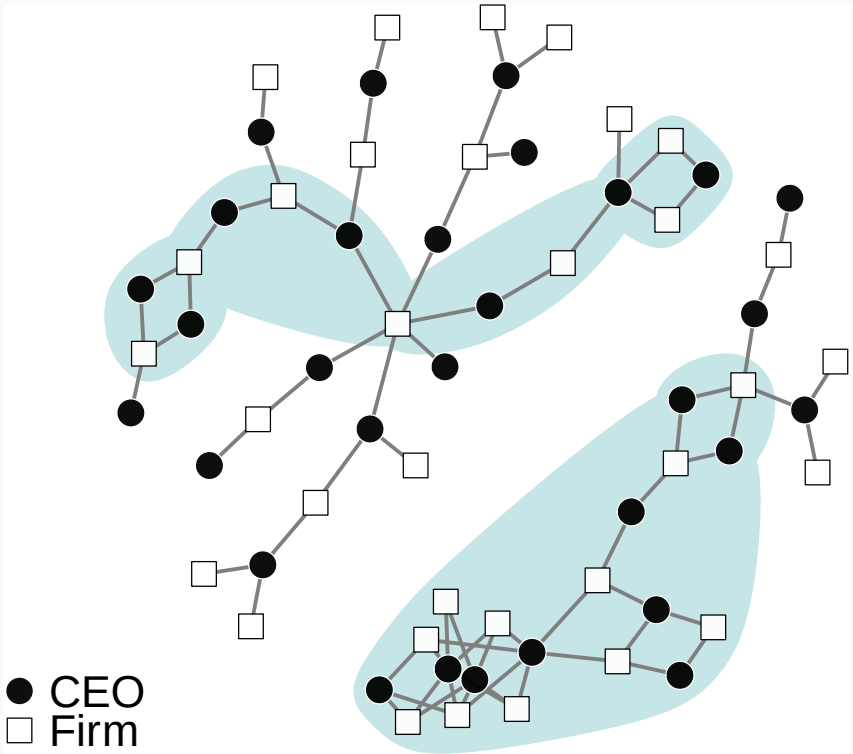


Figure 4: Graph of CEO–Firm Connections

Link with Graph Theory

- Solving a two–way fixed effects problem is exactly the same problem as an important problem in graph theory (solving $\mathbf{Lx} = \mathbf{b}$ where $\mathbf{L}$ is a Laplacian matrix)
- Spielman & Teng (2004), Kelner et al (2013):
    - Laplacian systems can now be solved in nearly–linear time, instead of in $O(n^{2.36})$!
    - This is a fundamental breakthrough in graph theory and numerical optimization, and we can apply it to solve our model
- Can also apply other insights from graph theory (e.g. regressions robust to changes in network topology)

Link with Graph Theory

However:

- Solver has a very complex implementation
- Suffers from cache locality problems (Hoske et al 2015, Boman et al 2016)
- Partially implemented with the `prune` option

Benchmarks

Benchmark

- Does the estimator converge?
- Is it feasible with large or very large datasets?
- What estimator/command should I use?

Benchmark - Datasets

- Until now most benchmarks have been performed with synthetic datasets
- Very poor match to real datasets (which often cannot be shared)
- New collection of synthetic and anonymized datasets: employer–employee (Portugal), borrower–bank (Peru), student–teacher (US), patent citations, open source contributions, directors and firms, etc.

Benchmark - Illustration

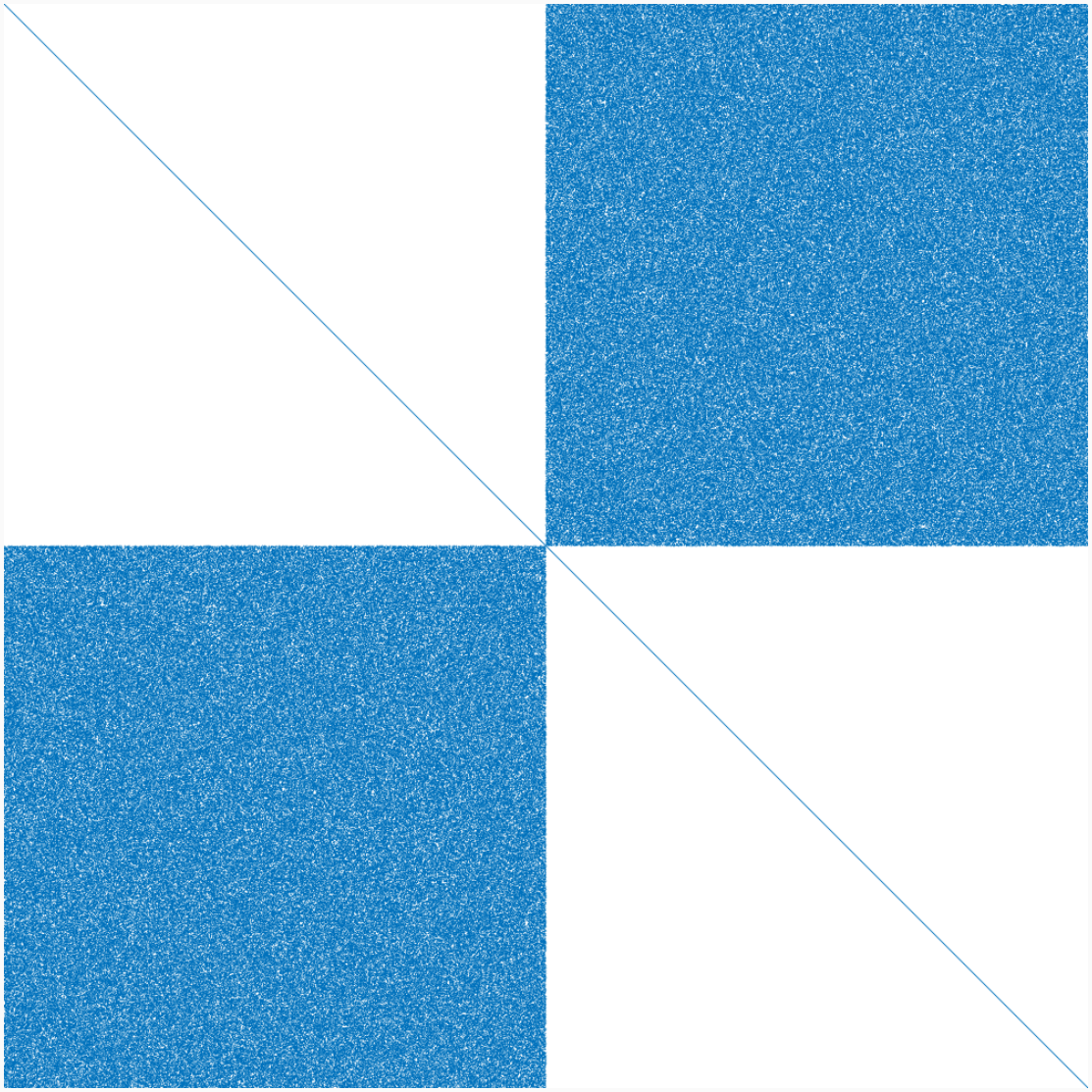


Figure 5: Synthetic Dataset (Gaure 2013)

Benchmark - Illustration

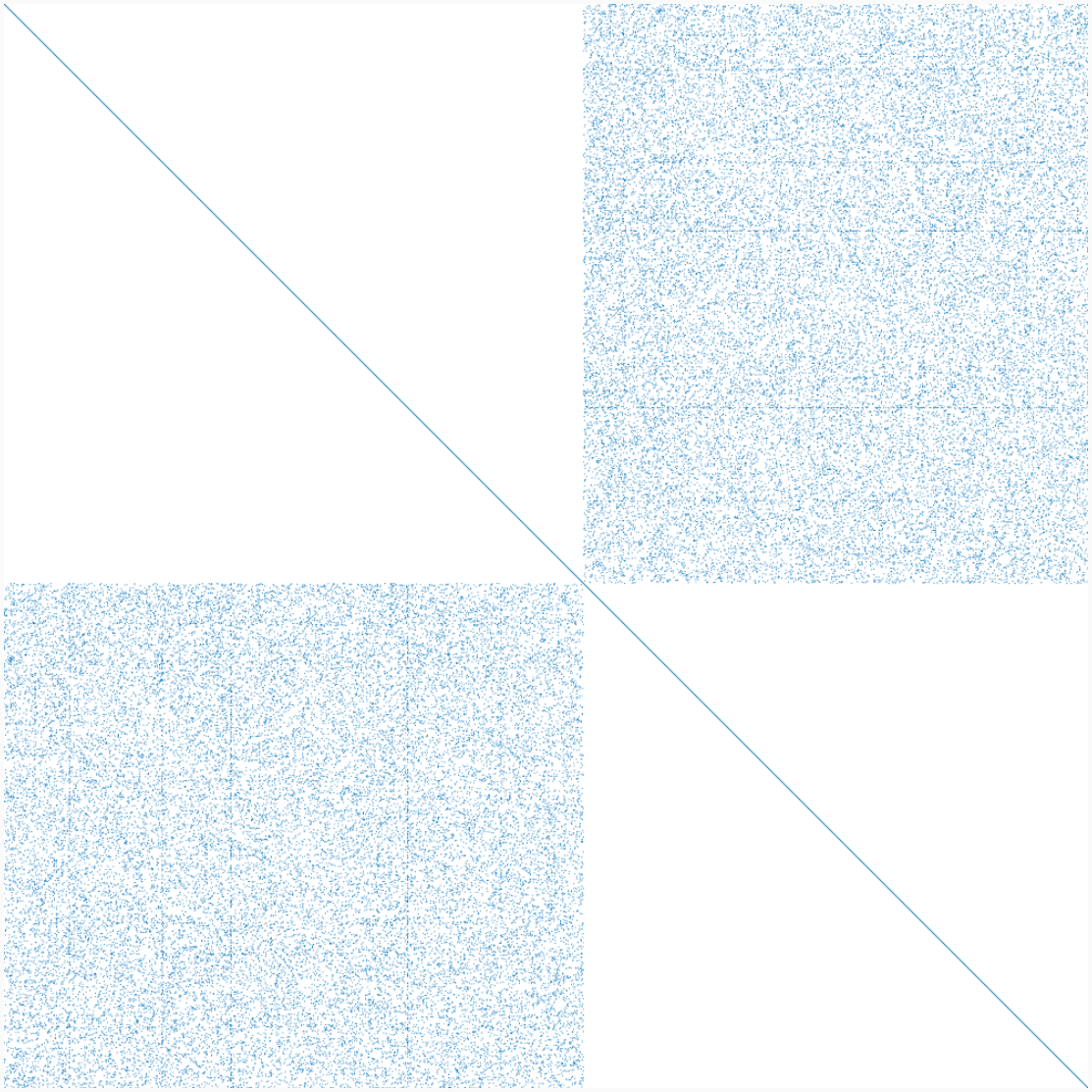


Figure 6: Github contributors and projects

Benchmark - Illustration

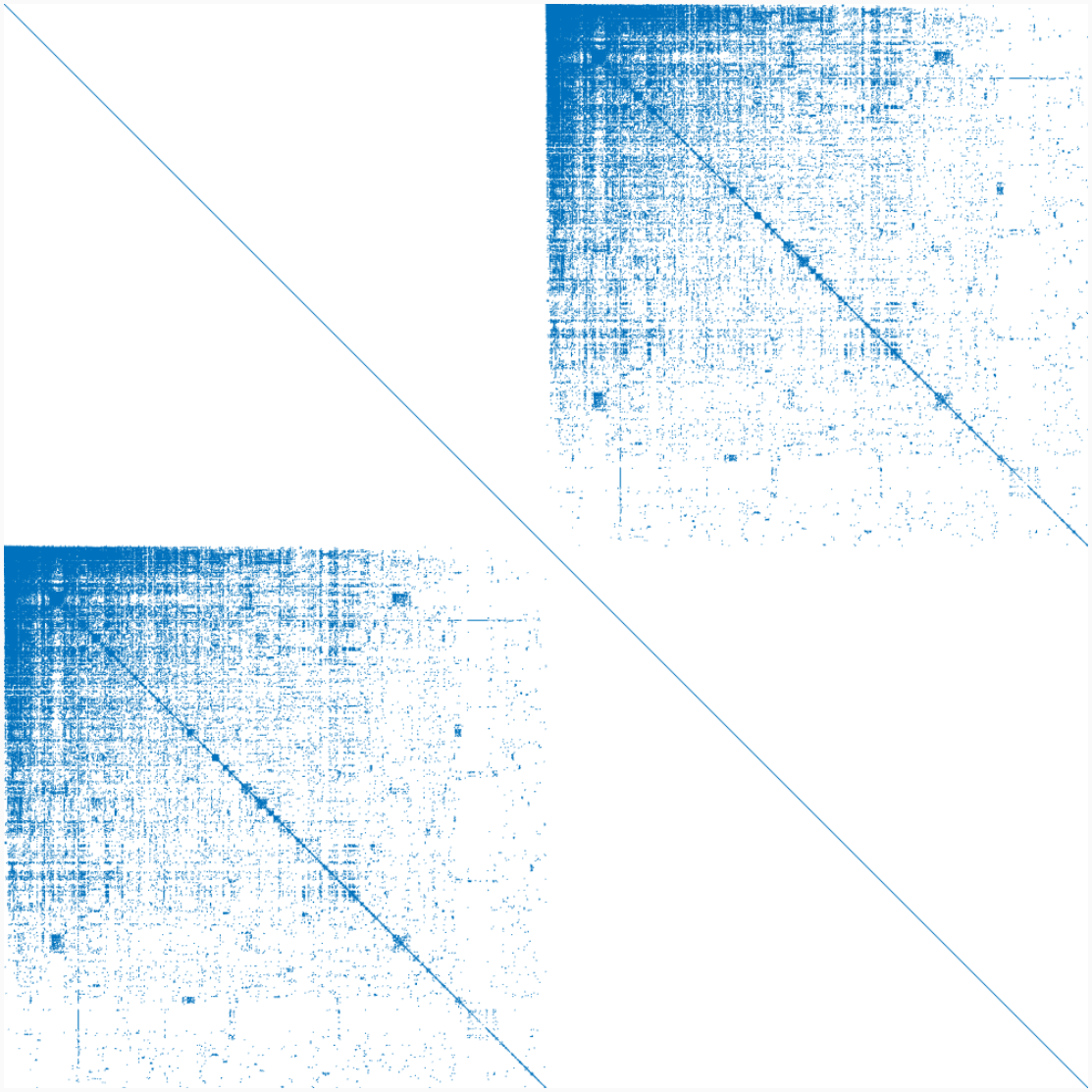


Figure 7: Enron Email Dataset (email senders and receivers)

Benchmark - Performance of Different `reghdfe` variants

	MAP-Aitken (Guim. 2012)	MAP-SD (Gaure 2013)	MAP-CG-Sym (Correia 2016)	MAP+Prune (Correia 2017)	LSMR (Gomez 2016)
Synthetic-complete	2.3	2.5	3.5	7.5	2.8
Synthetic-unif-easy	7.4	5.0	4.9	16.8	10.4
Synthetic-unif-hard	19.0	21.8	16.1	22.7	32.1
Synthetic-unif-harder	83.5	49.8	47.2	50.0	124.2
Synthetic-assortative	320.6	108.7	101.8	73.6	206.6
Credit	15.1	20.5	14.2	28.2	17.3
Enron	51.4	38.1	29.7	31.5	51.0
Schools	221.5	79.7	61.7	116.6	132.0
Github	268.7	110.1	114.4	127.1	169.6
Workers	484.1	146.0	169.6	646.1	356.3
Patents	574.2	191.7	188.5	172.3	490.6
Directors	688.8	215.6	258.6	240.7	894.1

Figure 8: Comparison of `reghdfe` variants across similar–sized datasets

Benchmark - Estimator Comparison

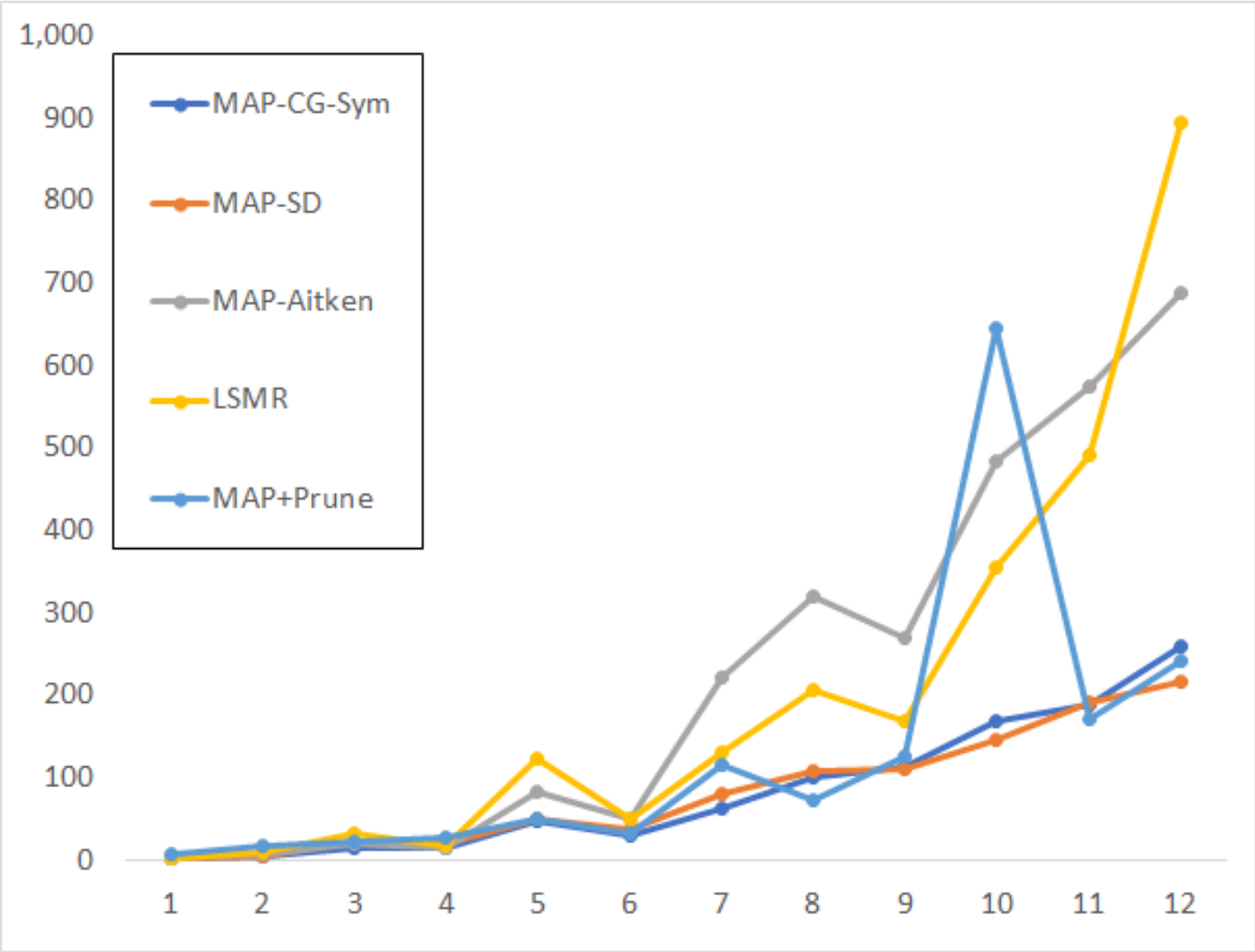
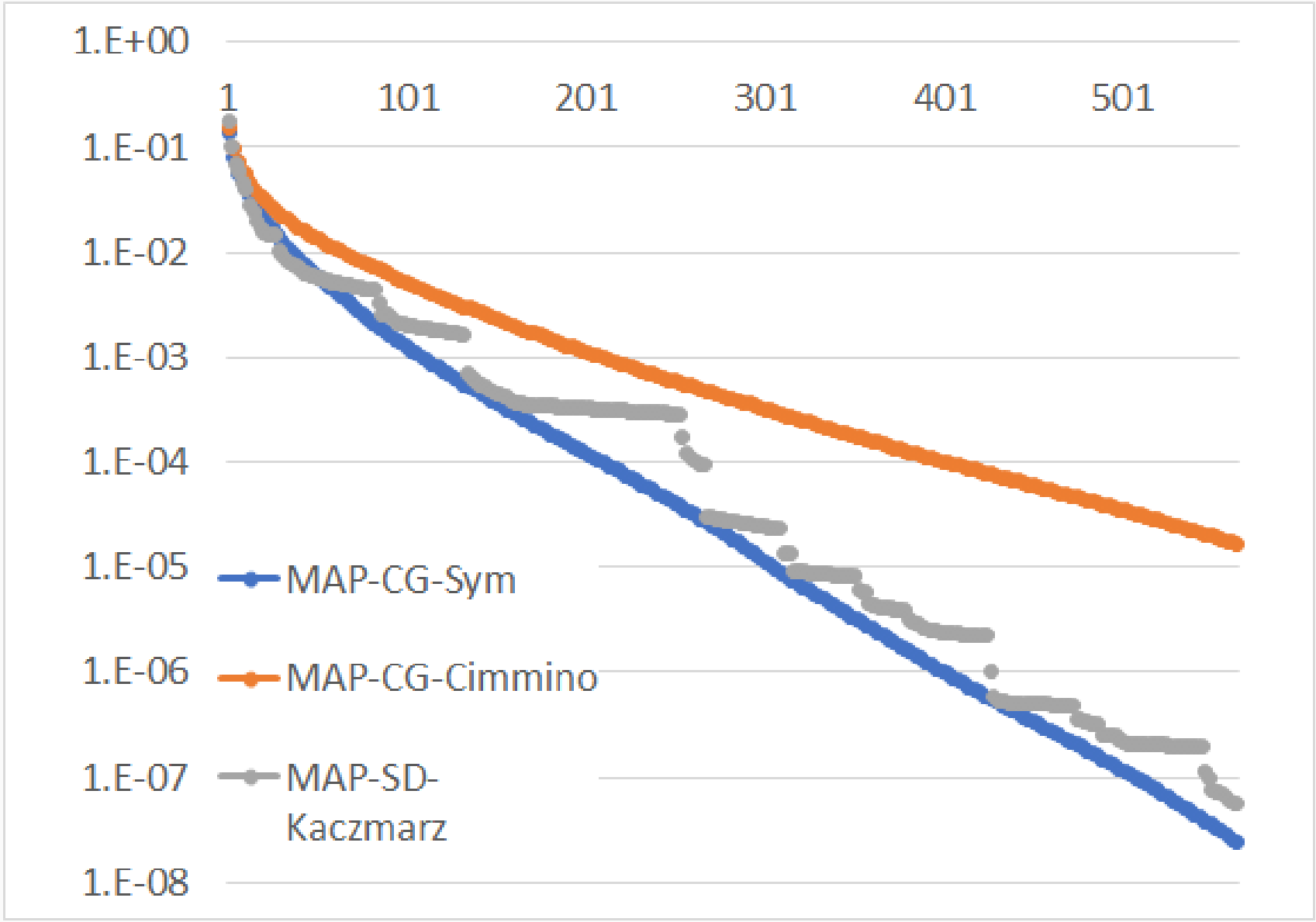


Figure 9: Comparison of `reghdfe` variants

Benchmark - Estimator Comparison



Benchmark - Estimator Comparison

Command	Synthetic Uniform	Synthetic Assortative	Synthetic Assortative	Github
reghdfe	2.0	0.4	71.5	33.0
fereg	3.0	34.5		
twfe	9.2	0.4	94.6	44.2
a2reg	11.4	1.1	122.1	
felsdvreg	29.2	3.6		
regxfe	36.2	66.7		
reg2hdfe	57.2	112.4		
Obs.	500,000	10,000	500,000	548,843

Figure 11: `reghdfe` and competing estimators

Warnings

Warnings - General Considerations

- Increase in sampling error:
    - Adding firm FEs losses between–firm variation, and so on. If most information is between units, there is not much left and standard errors will be large.
- Variance–covariance matrices are hard to compute:
    - Degrees-of-freedom cannot be easily computed and can be affected by singleton groups that bias the variance matrix.
- R2:
    - In models with many FEs, the R2 reflects the contribution of the FEs, not the main regressors. Solution: use within-R2

Warnings - FEs as the Object of Interest

- Many important papers have used the estimated fixed effects as input for subsequent regressions. EG: Bertrand-Schoar (2004)
- This is wrong:
  - Every FE estimate is inconsistent
  - F tests have an unknown distribution (Fee et al 2013)
  - R^2s are also biased
  - Correlations are downward biased (Gaure 2012)
  - Histograms and even placebo tests can be wildly misleading
  - WiP solution: Lamadon et al (2016), Arora et al (2017)

Warnings - Standard Errors are Inconsistent

Prologue:

- Singletons: $N/(N-K)$ vs $(N+M)/(N-K+M)$
- "Clone data"

```
reghdfe y x, absorb(firm) robust
expand 10, gen(id)
reghdfe y x, absorb(firm) cluster(id)
```

- Regressions at the state level vs. at county/zipcode level, if the data is almost unchanged between counties within a state

Warnings - Standard Errors are Inconsistent

Cattaneo, Jansson and Newey (2017):

"*if the number of included covariates are allowed to grow as fast as the sample size ... all of the usual versions of Eicker-White heteroskedasticity consistent standard error estimators for linear models are inconsistent*"

Warnings - Standard Errors are Inconsistent

- Solutions:
  - Compute $\mathbf{M} \otimes \mathbf{M}$ matrix; not possible with many FEs
  - Apply Verdier (2017) correction; extremely large computational costs (200 nodes in his paper)
  - Apply Wild Bootstrap; requires $K^{\delta}/N \to 0$ for $\delta > 1$
- Interpretation:
  - Need to understand how K grows with the dataset.
  - EG: adding interconnected obs. means we can apply wild bootstrap; else apply Verdier's approach

Thank you!