Branching fixed effects: A proposal for communicating uncertainty

Patrick Kline
UC Berkeley

ESWC 2025
Seoul, Korea

Introduction

Two-way fixed effects methods widely used to summarize complex causal relationships in administrative data

- Firm wage effects (Kline, 2024)
- Teacher value added (Angrist, Walters, Hull, 2023)
- Location effects (Chyn and Katz, 2021)

Data confidentiality / complexity leads to reliance on published fixed effects estimates for secondary analysis

Growing reliance on Other Peoples' Projections (OPP)

THE
QUARTERLY JOURNAL
OF ECONOMICS

Vol. 136 2021 Issue 2

LABOR IN THE BOARDROOM*

Simon Jäger
Benjamin Schoefer
Jörg Heining

We estimate the wage effects of shared governance, or codetermination, in the form of a mandate of one-third of corporate board seats going to worker representatives. We study a reform in Germany that abruptly abolished this mandate for stock corporations incorporated after August 1994, while it locked the mandate for the slightly older cohorts. Our research design compares firm cohorts

Can You Move to Opportunity? Evidence from the Great Migration†

By Ellora Derenoncourt*

This paper shows that racial composition shocks during the Great Migration (1940–1970) reduced the gains from growing up in the northern United States for Black families and can explain 27 percent of the region's racial upward mobility gap today. I identify northern Black share increases by interacting pre-1940 Black migrants' location choices with predicted southern county out-migration. Locational changes, not negative selection of families, explain lower upward mobility, with persistent segregation and increased crime and policing as plausible mechanisms. The case of the Great Migration provides a more nuanced view of moving to opportunity when destination reactions are taken into account. (*JEL* H75, H76, J15, J62, K42, N32, R23)

Econometrica, Vol. 90, No. 6 (November, 2022), 2567–2602

ROBUST EMPIRICAL BAYES CONFIDENCE INTERVALS

Timothy B. Armstrong
Department of Economics, University of Southern California

Michal Kolesár
Department of Economics, Princeton University

Mikkel Plagborg-Møller
Department of Economics, Princeton University

We construct robust empirical Bayes confidence intervals (EBCIs) in a normal means problem. The intervals are centered at the usual linear empirical Bayes estimator, but use a critical value accounting for shrinkage. Parametric EBCIs that assume a normal distribution for the means (Morris (1983b)) may substantially undercover when this assumption is violated. In contrast, our EBCIs control coverage regardless of the means distribution, while remaining close in length to the parametric EBCIs when the means are indeed Gaussian. If the means are treated as fixed, our EBCIs have an average coverage guarantee: the coverage probability is at least $1-\alpha$ on average across the n EBCIs for each of the means. Our empirical application considers the effects of U.S. neighborhoods on intergenerational mobility.

Keywords: Average coverage, empirical Bayes, confidence interval, shrinkage.

Predictive effects of teachers and schools on test scores, college attendance, and earnings

Gary E. Chamberlain[1]

Department of Economics, Harvard University, Cambridge, MA 02138

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2011.

Contributed by Gary E. Chamberlain, August 20, 2013 (sent for review April 18, 2013)

I studied predictive effects of teachers and schools on test scores in fourth through eighth grade and outcomes later in life such as college attendance and earnings. For example, predict the fraction of a classroom attending college at age 20 given the test score for a different classroom in the same school with the same teacher and given the test score for a classroom in the same school with a different teacher. I would like to have predictive effects that condition on averages over many classrooms, with and without the same teacher. I set up a factor model that, under certain assumptions, makes this feasible. Administrative school district data in combination with tax data were used to calculate estimates and do inference.

education production function | unmeasured inputs | teacher effects

not find a strong role for measured characteristics of teache as teacher experience, education, and test scores of teacher determination of academic achievement of students. Th a strong role for measured characteristics motivates intere measured characteristics of teachers that have a causal academic achievement. Related literature on estimating effects on test scores includes refs. 2–10. A typical findi a 1-SD increase in the teacher factor corresponds to an in individual scores on the order of 0.1, where the units are Sl distribution of scores for individual students.

In the Tennessee Student/Teacher Achievement Rati iment, known as Project STAR, children entering kind were randomly assigned to class types, which were r assigned to teachers. The random assignment was withi

Assessing uncertainty in OPP

Assumptions underlying published standard errors often dubious

- Usual "robust" standard errors inconsistent in high-dimensional models (Cattaneo, Jansson, Newey, 2018)
- Bootstrap likewise inconsistent (El Karoui and Purdom, 2018)

Generally need full covariance matrix for second-step inference

But challenging to release correlation matrix of estimates

- 1M firm effects involve $\sim$500B pairwise correlations
- Stored as floats, requires $\sim$1.82TB of storage!

Proposal: break estimates into independent "branches"

Branches are disjoint subsamples in which model can be estimated

- Each branch (ostensibly) identifies the same parameter
- Branch-specific estimates are mutually independent
- Enables transparent uncertainty quantification and shrinkage

Precedents

- "Replicates" in survey analysis (used in CPI / CPS / ACS to compute variance)
- Random sample-splitting in worker-firm literature (Sorkin, 2018; Card, Rothstein, Yi, 2024)

Technical contribution: computing maximal branchings

Goal: partition the microdata into as many branches as possible

Constraint: OLS design matrix must be full rank in each branch

Going to consider special case of "AKM model," where estimability is known to depend on connectedness (Abowd, Creecy, Kramarz, 2002)

- Minimally connected graph is a *spanning tree*
- Enumerate trees using results from graph literature on "tree packing" (Nash-Williams,1961; Tutte, 1961; Roskind and Tarjan, 1985)
- Assign one tree to each branch. Last branch gets "leftovers."

The blessings of branches

Preliminaries

Suppose we are interested in parameter vector $\psi \in \mathbb{R}^J$

- Linear FE estimator $\hat{\psi}$ is unbiased $\left(\mathbb{E}\left[\hat{\psi}\right] = \psi\right)$
- Given M branches, can construct M mutually independent estimates: $\left\{\hat{\psi}_b\right\}_{b=1}^{M}$
- Under linear model assumptions, each branch-specific estimate is unbiased:
$$\mathbb{E}\left[\hat{\psi}_b\right] = \mathbb{E}\left[\hat{\psi}\right] = \psi$$

Linear decomposition

The following linear decomposition property will be shown to hold:

$$\hat{\psi} = \sum_{b=1}^{M} \boldsymbol{C}_b \hat{\psi}_b \equiv \sum_{b=1}^{M} \varphi_b,$$

where $\boldsymbol{C}_b$ is a known $J \times J$ matrix that obeys $\sum_{b=1}^{M} \boldsymbol{C}_b = \boldsymbol{I}$

- Each φ_b gives the contribution of branch b to full-sample $\hat{\psi}$
- Influence fn interpretation

Proposal: publish the pairs $\left\{\hat{\psi}_b, \varphi_b\right\}_{b=1}^{M}$

Putting branches to work: variance estimation

Since the branches are independent,

$$\mathbb{V}\left[\hat{\psi}\right] = \sum_{b=1}^{M} \mathbb{V}\left[\varphi_b\right] = \sum_{b=1}^{M} \boldsymbol{C}_b \mathbb{V}\left[\hat{\psi}_b\right] \boldsymbol{C}_b' \equiv \Sigma.$$

For any two branches b and ℓ,

$$\mathbb{E}\left[\left(\varphi_b - \varphi_\ell\right)\left(\varphi_b - \varphi_\ell\right)'\right] = \underbrace{\boldsymbol{C}_b \mathbb{V}\left[\hat{\psi}_b\right] \boldsymbol{C}_b'}_{=\mathbb{V}[\varphi_b]} + \underbrace{\boldsymbol{C}_\ell \mathbb{V}\left[\hat{\psi}_\ell\right] \boldsymbol{C}_\ell'}_{=\mathbb{V}[\varphi_\ell]}.$$

Putting branches to work: variance estimation

Summing across all $\binom{M}{2}$ pairs of branches, yields the unbiased variance estimator

$$\hat{\Sigma} = \frac{1}{M-1} \sum_{b=1}^{M} \sum_{\ell<b} \left(\varphi_b - \varphi_\ell\right)\left(\varphi_b - \varphi_\ell\right)'$$

- Note: yields estimates of all $J\left(J-1\right)/2$ off-diagonal terms!
- While $\hat{\Sigma}$ is noisy, typically interested in low-dimensional quadratic forms $v'\Sigma v$ that can be estimated consistently.
- High-level conditions for $v'\hat{\Sigma}v \xrightarrow{p} v'\Sigma v$ or can estimate $\mathbb{V}\left[v'\hat{\Sigma}v\right]$ when $M \geq 4$.

Putting branches to work: moment estimation

Let $\odot$ denote the element-wise product operator. For any two branches b and ℓ,

$$\mathbb{E}\left[\hat{\psi}_b \odot \hat{\psi}_\ell\right] = \psi \odot \psi$$

- Averaging across all $\binom{M}{2}$ distinct pairs of branches yields the second moment estimator:

$$\frac{1}{J}\mathbf{1}'\left[\frac{2}{M\left(M-1\right)}\sum_{b=1}^{M}\sum_{\ell<b}\hat{\psi}_b \odot \hat{\psi}_\ell\right]$$

- Higher moments can be estimated with higher-order products

Putting branches to work: shrinkage

Let $\hat{\psi}_{bj}$ denote the j'th entry of $\hat{\psi}_b$ and ψ_j the j'th entry of ψ.

Now consider random effects model:

$$\hat{\psi}_b \mid \psi \sim F_b, \quad \psi_j \overset{i.i.d}{\sim} G,$$

where $\mathbb{E}_{F_b}\left[\hat{\psi}_b\right] = \psi$. By iterated expectations (Krutchkoff, 1967),

$$\begin{aligned} \mathbb{E}\left[\hat{\psi}_{2j} \mid \hat{\psi}_{1j}\right] &= \mathbb{E}\left[\mathbb{E}\left[\hat{\psi}_{2j} \mid \psi_j, \hat{\psi}_{1j}\right] \mid \hat{\psi}_{1j}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\hat{\psi}_{2j} \mid \psi_j\right] \mid \hat{\psi}_{1j}\right] \\ &= \mathbb{E}\left[\psi_j \mid \hat{\psi}_{1j}\right] \end{aligned}$$

$\Rightarrow$ a regression of estimates in one branch against the other gives a best predictor regardless of noise distribution F_b!

AKM basics

The AKM model

TWFE model for log wages of worker $i \in \{1, 2, \ldots, N\} \equiv [N]$ in period $t \in \{1, 2\}$ (Abowd, Kramarz, Margolis, 1999):

$$y_{it} = \underbrace{\alpha_i}_{\text{worker FE}} + \underbrace{\psi_{\mathbf{j}(i,t)}}_{\text{firm FE}} + \underbrace{\varepsilon_{it}}_{\text{noise}}$$

Function $\mathbf{j} : [N] \times \{1, 2\} \to \{1, 2, \ldots, J\} \equiv [J]$ gives the firm employing worker i in period t.

Let $\mathcal{P}$ denote set of O/D firm pairs that workers move between:

$$\left\{(o, d) \in [J]^2 : o \neq d,\ \mathbf{j}(i, 1) = o,\ \mathbf{j}(i, 2) = d \text{ for some } i \in [N]\right\}$$

First differenced representation

We are interested in the firm effects $\psi = (\psi_1, \dots, \psi_J)'$. Eliminate worker FE via first differencing and write in matrix notation:

$$\boldsymbol{y}_2 - \boldsymbol{y}_1 = (\boldsymbol{F}_2 - \boldsymbol{F}_1)\,\psi + \boldsymbol{\varepsilon}_2 - \boldsymbol{\varepsilon}_1 \equiv \boldsymbol{P}\boldsymbol{B}'\psi + \boldsymbol{u} \qquad (1)$$

- $\boldsymbol{P}$ is $N \times |\mathcal{P}|$ matrix of origin-destination pair indicators.
- $\boldsymbol{B}$ is $J \times |\mathcal{P}|$ "incidence matrix," each column of which is sparse with a single $+1$ (destination) and a single -1 (origin).

Exogeneity:

$$\mathbb{E}\left[\boldsymbol{u} \mid \boldsymbol{P}, \boldsymbol{B}\right] = 0$$

Collapsing to mean wage changes

The diagonal matrix $\boldsymbol{W} = \boldsymbol{P}'\boldsymbol{P}$ gives number of moves and

$$\hat{\Delta} = \left(\boldsymbol{P}'\boldsymbol{P}\right)^{-1} \boldsymbol{P}' \left(\boldsymbol{y}_2 - \boldsymbol{y}_1\right)$$

is vector of mean wage changes for O/D pairs.

Rearranging (1) yields

$$\boldsymbol{B}\boldsymbol{W}\hat{\Delta} = \boldsymbol{B}\boldsymbol{W}\boldsymbol{B}'\psi + \boldsymbol{B}\boldsymbol{P}'\boldsymbol{u}$$

Key moment restriction:

$$\mathbb{E}\left[\boldsymbol{B}\boldsymbol{W}\hat{\Delta}\right] = \boldsymbol{B}\boldsymbol{W}\boldsymbol{B}'\psi \equiv \boldsymbol{L}\psi$$

$\boldsymbol{L}$ is the (weighted) *Laplacian* of worker mobility graph.

Graphical representation

The mobility graph

Define the *mobility graph* $G = (V, E)$

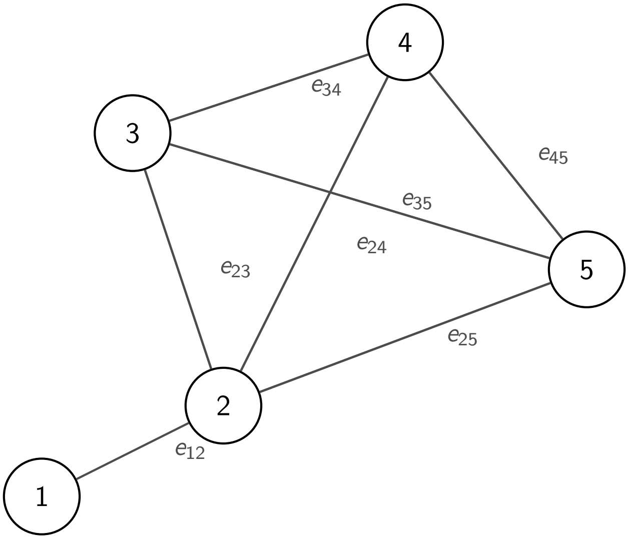
- Vertex set $V = [J]$ corresponds to firms
- E is set of *edges* defined by worker moves:

$$\{(s, v) \in [J]^2 : s \neq v,\ \mathbf{j}(i, 1) = s,\ \mathbf{j}(i, 2) = v \text{ for some } i \in [N]\}$$

Treat edges as *undirected*:

$$e_{sv} = e_{vs}$$

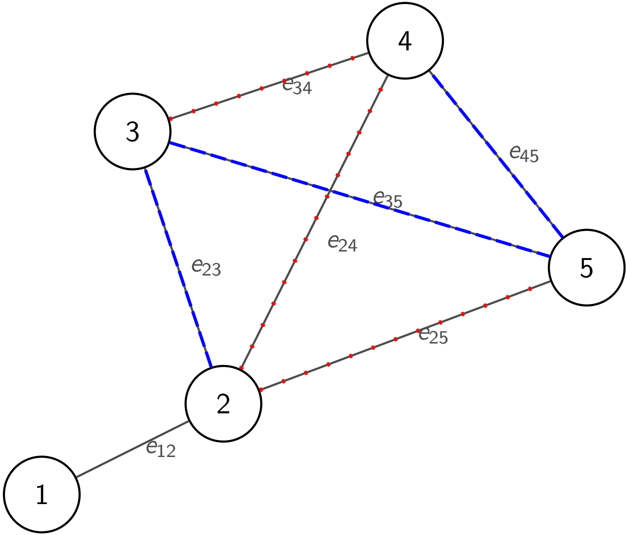
A mobility graph with $J = 5$ and $|E| = 7$



Terms of art (#IYKYK)

- A *path* is a sequence of edges that join a set of vertices with no repeat edges or firms: e.g., $\{e_{12}, e_{23}, e_{34}\}$
- A graph is *connected* if there is a path from any firm to any other firm
- A *tree* is a connected graph in which there is a unique path between any two firms
- A *spanning tree* is any subset of a connected graph that contains all firms and is a tree

Two (edge-disjoint) spanning trees of subgraph



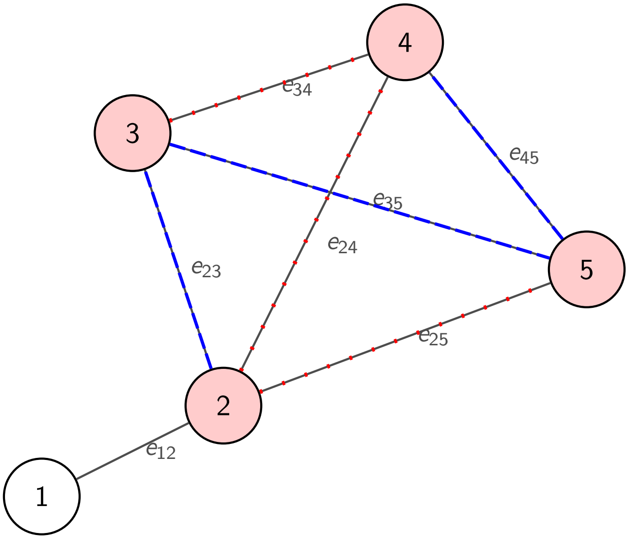
Connectivity

The *edge connectivity* $\lambda(G)$ of a graph is the number of edges that need to be removed from the graph for it to become disconnected:

$$\lambda(G) = \min\Big\{|S| : S \subseteq E, \text{ and } (V, E \setminus S) \text{ is disconnected}\Big\}$$

- A graph is *k-edge connected* if $\lambda(G) \geq k$
- A *k-edge connected component* (k-ECC) is a subgraph that is k-edge connected
- There exist k *edge-disjoint paths* between any two firms in a k-ECC (Menger, 1927)

This graph has a single 3-ECC



Back to OLS

Recall that AKM yields the normal equations

$$\boldsymbol{BW}\hat{\Delta} = \boldsymbol{L}\psi$$

Theorem (Kirchhoff, 1847; Abowd, Creecy, Kramarz, 2002)

If the mobility graph is connected, then $\boldsymbol{L}$ *has rank* $J-1$ *and any minor of* $\boldsymbol{L}$ *will have full rank.*

Identification: normalize $\psi_1 = 0$

- Define $\psi_{(1)} = (\psi_2, \ldots, \psi_J)'$
- Vector $\psi_{(1)}$ measures expected wage changes associated with leaving firm $\#1$ for each destination.

Normalization and estimation

Define $\boldsymbol{B}_{(1)}$ as $\boldsymbol{B}$ with first row deleted and

$$\boldsymbol{L}_{(1)} = \boldsymbol{B}_{(1)} \boldsymbol{W} \boldsymbol{B}_{(1)}$$

OLS estimator of $\psi_{(1)}$ is:

$$\hat{\psi}_{(1)} = \boldsymbol{L}_{(1)}^{-1} \boldsymbol{B}_{(1)} \boldsymbol{W} \hat{\Delta}$$

Firm effects a linear combination of O/D wage changes $\hat{\Delta}$

- When moves present in both directions between firms j and k, only net change between pair matters
- For today, suppose moves in single direction between each firm pair (general case in the paper)

Building branches

Special case: tree structure

When $\boldsymbol{B}_{(1)}$ is invertible square matrix, we get simplification

$$\hat{\psi}_{(1)} = \left(\boldsymbol{B}'_{(1)}\right)^{-1} \hat{\Delta}$$

Mobility pattern is a *tree* (Kline, 2024)

- Unique path between any pair of firms
- Each $\hat{\psi}_j$ equals sum of (oriented) entries in $\hat{\Delta}$ along path from firm 1 to firm j
- Deleting moves between any pair of firms would render model under-identified
- Weights irrelevant because "just-identified"

Branches from trees

WLOG, can decompose incidence matrix of connected graph as

$$\boldsymbol{B}_{(1)} = \left[\boldsymbol{T}_{(1),1}, \ldots, \boldsymbol{T}_{(1),M-1}, \boldsymbol{T}^{+}_{(1),M}\right]$$

- $\boldsymbol{T}_{(1),1}, \ldots, \boldsymbol{T}_{(1),M}$ capture edge-disjoint spanning trees
- Usually some "leftover" edges that are not connected
- $\boldsymbol{T}^{+}_{(1),M}$ captures the last tree + leftovers

Likewise, partition wage changes and weights as

$$\hat{\Delta} = \left(\hat{\Delta}'_1, \ldots, \hat{\Delta}'_M\right)', \quad \mathbf{W} = diag\left(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_M\right)$$

Branch-specific OLS estimates

We can now define branch-specific estimates

$$\hat{\psi}_{(1),b} = \begin{cases} \left(\boldsymbol{T}_{(1),b} \boldsymbol{T}'_{(1),b} \right)^{-1} \boldsymbol{T}_{(1),b} \hat{\Delta}_b & \text{if } b < M \\ \left(\boldsymbol{T}^+_{(1),M} \boldsymbol{W}_M \left(\boldsymbol{T}^+_{(1),M} \right)' \right)^{-1} \boldsymbol{T}^+_{(1),M} \boldsymbol{W}_M \hat{\Delta}_M & \text{if } b = M \end{cases}$$

Recall that full-sample is related to branch-specific estimates by

$$\hat{\psi}_{(1)} = \sum_{b=1}^{M} \boldsymbol{C}_{(1),b} \hat{\psi}_{(1),b} = \sum_{b=1}^{M} \varphi_{(1),b}$$

Combination matrices are ratio of branch to full-sample Laplacian

$$\boldsymbol{C}_{(1),b} = \begin{cases} \boldsymbol{L}^{-1}_{(1)} \boldsymbol{T}_{(1),b} \boldsymbol{W}_b \boldsymbol{T}'_{(1),b} & \text{if } b < M \\ \boldsymbol{L}^{-1}_{(1)} \boldsymbol{T}^+_{(1),M} \boldsymbol{W}_M \left(\boldsymbol{T}^+_{(1),M} \right)' & \text{if } b = M \end{cases}$$

Branch contributions easy to compute

Structure of $\boldsymbol{C}_{(1),b}$ implies branch contributions are

$$\varphi_{(1),b} = \begin{cases} \boldsymbol{L}_{(1)}^{-1} \boldsymbol{T}_{(1),b} \boldsymbol{W}_b \hat{\Delta}_b & \text{if } b < M \\ \boldsymbol{L}_{(1)}^{-1} \boldsymbol{T}_{(1),M}^{+} \boldsymbol{W}_M \hat{\Delta}_M & \text{if } b = M \end{cases}$$

- Each branch contribution is a regression coefficient with wage changes as dep var
- No harder to compute than branch-specific estimates $\hat{\psi}_{(1),b}$

Finding the trees

Tree Packing 101

Connectivity $\lambda(G)$ provides bounds on # of edge-disjoint trees that can be "packed" into a graph

Theorem (Nash-Williams-Tutte, 1961)

The number $\tau(G) \in \mathbb{N}$ of spanning trees that can be packed into a connected graph $G = (V, E)$ obeys

$$\lfloor \lambda(G)/2 \rfloor \leq \tau(G) \leq \lambda(G),$$

where $\lfloor \cdot \rfloor$ denotes the floor operator.

Hence, a k-ECC must contain at least $\lfloor k/2 \rfloor$ edge-disjoint trees. But finding these trees is non-trivial...

“Greedy” extraction can miss trees

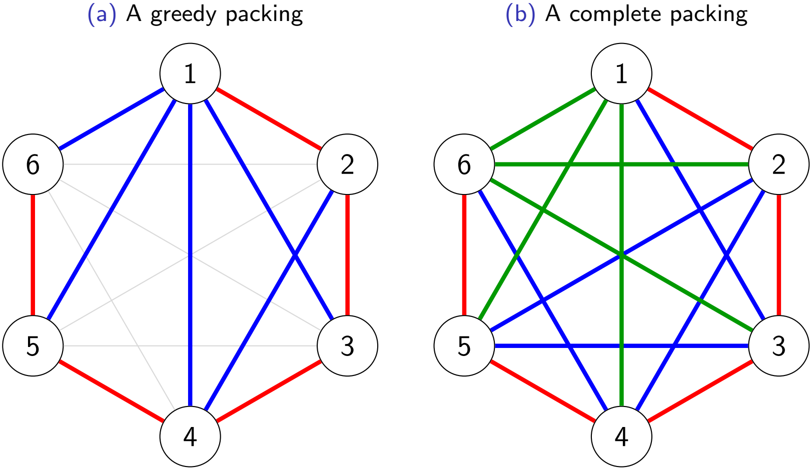


Figure: Two attempts to pack the same graph

An iterative solution

Tree packing is an ILP with high-dimensional constraints

$$\begin{aligned} \max_{x} \quad & \sum_{T \in \mathcal{T}} x_T \\ \text{s.t.} \quad & \sum_{T \ni e} x_T \leq 1 \quad \forall\, e \in E, \\ & x_T \in \{0, 1\} \quad \forall\, T \in \mathcal{T}, \end{aligned}$$

where $\mathcal{T}$ is set of all trees in graph. Infeasible to store constraints.

Roskind and Tarjan (1985) provide iterative algorithm guaranteed to pack k-ECC with M edge-disjoint spanning trees if $M \leq \tau(G)$

- Highly scalable: computational complexity $\mathcal{O}\left(M^2 J^2\right)$
- Rely on SageMath implementation

A "prune and pack" strategy

Prune and pack strategy:

1. Prune graph to k-ECC using Gusfield (1990) algorithm
2. Find all edge-disjoint trees via Roskind-Tarjan (1985)

Implement for different choices of k

- Useful to release results for multiple choices of k
- SageMath Jupyter notebook coming soon..

Results

Application: firm effects in Veneto, Italy

Revisit sample studied in Kline, Saggio, and Solvsten (2020): firm switchers with wage observations 1999 and 2001

Dimensions:

- 73,933 firms
- 197,572 movers
- 148,917 (undirected) edges

Notes:

- Only 1.3 movers per edge $\Rightarrow$ little scope to gain branches by packing directed trees
- Low degree: 2.7 movers per firm $\Rightarrow$ extremely low connectivity

Pruning drops small firms

k	1	2	3	4	5	6
Firms in k-core	73,933	41,093	21,570	11,145	5,682	3,128
Firms in k-ECC (J)	73,933	41,054	21,565	11,145	5,682	3,128
Edges in k-ECC ($\|E\|$)	148,917	116,026	80,561	51,824	31,677	19,796
Movers in k-ECC (N)	197,572	158,149	114,717	78,908	53,131	36,903
Spanning Trees (M)	1	1	2	3	3	4

Table: Network properties of k-ECCs

Pruning to 3-ECC yields second spanning tree

- Loses 61% of the firms but retains 58% of the movers
- Can mitigate firm loss by "filling in" graph using additional yrs of data

Random splits retain slightly more firms

Splits (M)	1	2	3	4
Number of firms				
25th Percentile	73,933	28,680	13,034	6,537
Median	73,933	28,737	13,077	6,576
75th Percentile	73,933	28,797	13,127	6,607
Overlap across simulations				
25th Percentile	73,933	22,745	9,328	4,482
Median	73,933	22,804	9,366	4,509
75th Percentile	73,933	22,861	9,406	4,536

Table: Firm effects estimable in each of M random splits (500 sims)

- Size of largest component stable
- But composition highly variable

Use case: firm size wage premium

Define covariate matrix $\boldsymbol{X} = [\mathbf{1}, \ln f]$, where f is firm size. We are interested in projection slope

$$\gamma = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\psi$$

Estimator is

$$\hat{\gamma} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\hat{\psi}$$

Variance is

$$\mathbb{V}\left[\hat{\gamma}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\Sigma\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Variance estimator is

$$\hat{\mathbb{V}}_{branch}\left[\hat{\gamma}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\hat{\Sigma}\boldsymbol{X} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

Branch-based standard errors 2-4x naive

k	1	2	3	4	5	6
Full-sample estimate ($\hat{\gamma}_{(1)}$)	0.0446	0.0329	0.0235	0.0199	0.0191	0.0209
Standard error ($\hat{\mathbb{V}}_{branch}\left[\hat{\gamma}_{(1)}\right]^{1/2}$)	-	-	0.0080	0.0045	0.0084	0.0091
Naive std err ($\hat{\mathbb{V}}_{HC0}\left[\hat{\gamma}_{(1)}\right]^{1/2}$)	0.0009	0.0010	0.0013	0.0018	0.0023	0.0032
Mean firm size	13.6	21.6	34.6	55.1	88.0	131.6

Table: Elasticity of firm wage effects with respect to firm size

- Naive t-stats: 7-20
- Branch-based t-stats: 2.3-2.9
- Some of this decrease likely reflects misspecification

Use case: size-weighted moments of firm effects

Let $\omega \in \mathbb{R}^{J}_{\geq 0}$ be vector of weights that sums to one $\left(\mathbf{1}'\omega = 1\right)$.

Plug-in estimator of the ℓ'th central weighted moment is

$$\hat{\mu}_{\ell,PI} = \omega' \left(\hat{\psi} - \omega'\hat{\psi}\mathbf{1}\right)^{\circ\ell},$$

where the $\circ\ell$ superscript denotes raising the entries in a vector to the ℓ'th power element-wise.

Use branches to construct the corresponding unbiased estimator:

$$\hat{\mu}_{\ell} = \binom{M}{\ell}^{-1} \sum_{(b_1,\ldots,b_\ell)\in\mathcal{B}} \omega' \left(\hat{\psi}_{b_1} - \omega'\hat{\psi}_{b_1}\mathbf{1}\right) \odot \cdots \odot \left(\hat{\psi}_{b_\ell} - \omega'\hat{\psi}_{b_\ell}\mathbf{1}\right)$$

Firm effects are skew left and heavy tailed

k	1	2	3	4	5	6
Std dev ($\sigma \equiv \sqrt{\mu_2}$)						
Plug-in	0.2215	0.1951	0.1841	0.1794	0.1751	0.1756
Branches	-	-	0.1591	0.1669	0.1917	0.1555
Skew (μ_3/σ^3)						
Plug-in	-0.8332	-0.7403	-0.8223	-0.8507	-0.7198	-0.6998
Branches	-	-	-	-1.217	-0.7911	-0.7368
Kurtosis (μ_4/σ^4)						
Plug-in	7.309	5.824	6.137	6.191	5.776	5.416
Branches	-	-	-	-	-	6.617

Table: Moments of firm effects (size-weighted)

- Mild upward bias in PI estimate of std dev
- Negligible bias in PI skew
- Mild downward bias in PI Kurtosis
- Miniscule std err on $\hat{\mu}_2$ in 6-ECC (z-score $\sim$30)

Shrinkage: non-linear relationship between branches

(a) $\mathbb{E}\left[\hat{\psi}_{2j} \mid \hat{\psi}_{1j}\right]$ (b) $\mathbb{E}\left[\hat{\psi}_{1j} \mid \hat{\psi}_{2j}\right]$

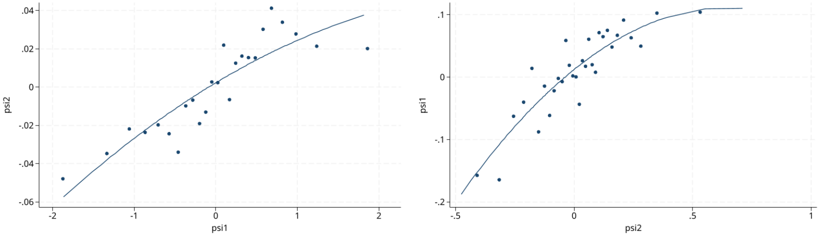


Figure: Binscatters of polynomial fit to opposite branch

Assessing forecast performance in the 6-ECC

Hold out $\hat{\psi}_4$ as forecast target and shrink $\left(\hat{\psi}_1, \hat{\psi}_2, \hat{\psi}_3\right)$ to construct predictor

- Naive predictor is average across branches:

$$\bar{\psi} = \left(\hat{\psi}_1 + \hat{\psi}_2 + \hat{\psi}_3\right)/3$$

- Forecast error is

$$MSE_{\text{naive}} = \left(\hat{\psi}_4 - \bar{\psi}\right)'\left(\hat{\psi}_4 - \bar{\psi}\right)/J$$

Shrinkage

Estimate cross-branch conditional expectations

$$\mathbb{E}\left[\hat{\psi}_{b_1 j} \mid \hat{\psi}_{b_2 j} = x_2, \hat{\psi}_{b_3 j} = x_3\right] \equiv m_{b_1}\left(x_2, x_3\right)$$

via B-spline series regression.

- Shrinkage predictor averages across branch pairs

$$\begin{aligned} \bar{m}\left(\hat{\psi}_{1j}, \hat{\psi}_{2j}, \hat{\psi}_{3j}\right) &= \frac{1}{3}\hat{m}_1\left(\hat{\psi}_{2j}, \hat{\psi}_{3j}\right) + \frac{1}{3}\hat{m}_2\left(\hat{\psi}_{1j}, \hat{\psi}_{3j}\right) \\ &+ \frac{1}{3}\hat{m}_3\left(\hat{\psi}_{1j}, \hat{\psi}_{2j}\right) \end{aligned}$$

- Contrast with Ignatiadus et al (2023) procedure (AURORA) designed for replicates.

Massive reduction in forecast error via shrinkage

	Naive $(\bar{\psi})$	Shrinkage $(\bar{m})$	AURORA
MSE	0.129	0.041	0.042

Table: Predicting $\hat{\psi}_4$ using $\left(\hat{\psi}_1, \hat{\psi}_2, \hat{\psi}_3\right)$

- Shrinkage lowers MSE by a factor of 3!
- AURORA fails to improve performance, perhaps because of heteroscedasticity across branches

Conclusion

Sharing estimates is good (Andrews and Shapiro, 2021; Donoho, 2024)

Sharing branches is even better

- Transparent uncertainty quantification
- Moment estimation
- Nonlinear shrinkage

Areas for future work

- Beyond AKM: branching more complex models
- Misspecification: separating noise from model error