Optimal Estimation of Two-Way Effects under Limited Mobility

Xu Cheng Sheng Chao Ho Frank Schorfheide

University of Pennsylvania

FRB Philadelphia Seminar
September 2023

Introduction

- Matched Data / Interaction-Based Model
  - student and teacher;
  - employee and employer;
  - patient and care provider
- Agent Specific Parameters for Unobserved Heterogeneity
  - modeled by two-way effects, outcome depends on pair, i.e., $\alpha_i + \beta_j$;
  - allow for assortative matching;
  - condition on the matching network.
- Running Example: estimation of teacher-value added.

Introduction: Scarce Information is a Challenge

- Limited Observations Per Agent
  - Teachers: limited class size;
  - Students: observations for only a few years
- Limited Mobility Across Agents
  - Identification of teacher value-added is based on students moving from one teacher to another.
  - Limited mobility can be represented as weak connectivity in a bipartite graph connecting teachers and students.

Contribution: A New Estimator Robust to Weak Connectivity

- Empirical Bayes (shrinkage) estimator for two-way effects.
- Adaptive to level of mobility/connectivity through hyperparameter estimation based on unbiased risk criterion.
- We establish asymptotic optimality within a class of estimator.
- Monte Carlo study and empirical application: estimation of teacher value-added based on a matched student-teacher data set.

Literature I

- OLS Estimator for Two-Way Effects is Widely Applied
  - employer-employee: Abowd, Kramarz, and Margolis (1999); Card, Heining, and Klein (2013); etc.
  - student-teacher, school-teacher: Clotfelter, Ladd, and Vigdor (2007); Jackson, Rockoff, and Staiger (2014); Mansfield (2015); etc.
  - demand and supply of health care: Finkelstein, Gentzkow, and Williams (2016).

- Issues with OLS Estimator for Two-Way Effects
  - Jochmans and Weidner (2019): finite-sample variance and large-sample consistency depend on connectivity measures of the network.
  - Verdier (2020) studies homogeneous regression coefficient estimation.

Literature II

- Shrinkage Estimation and Empirical Bayes Methods
  - James and Stein (1961); Lindley (1962); Stein (1962); Efron and Morris (1972, 1973); Stein (1981); etc.
  - Xie, Kou, and Brown (2012, 2016); Brown, Mukherjee, and Weinstein (2018); Kwon (2021); etc.
  - Robbins (1951, 1956); Brown and Greenshtein (2009); Koenker and Mizera (2014); Gu and Koenker (2017a,b); Liu, Moon, and Schorfheide (2020); etc.

- Existing Shrinkage Estimation for Teacher Value Added
  - one-way effect: Kane, Rockoff, and Staiger (2006); Kane and Staiger (2008); Chetty, Friedman, and Rockoff (2014); Gilraine, Gu, and McMillan (2020); Kwon (2021); etc.

Econometric Model

- Model, e.g., for test score:

$$y_{it} = \alpha_i + \beta_{j(i,t)} + x_{it}'\gamma + u_{it},$$

  - student $i \in \mathcal{S} = \{1, \ldots, r\}$, r is “rows”;
  - teacher $j(i,t) \in \mathcal{T} = \{1, \ldots, c\}$ of student i in time t, c is “columns”;
  - $u_{it} \mid j(\cdot), \alpha_{1:r}, \beta_{1:c} \sim_{iid} (0, \sigma^2)$;
  - for presentation, we assume there are no covariates x_{it}.

- In this talk: estimate $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_c)'$, e.g., teacher value added.
- Normalization: $1_c'\boldsymbol{\beta} = 0$.

Econometric Model

Vector Notation

$$\begin{aligned}\boldsymbol{Y} &= B_1\boldsymbol{\alpha} + B_2\boldsymbol{\beta} + \boldsymbol{U}, \quad \boldsymbol{U} \mid (B, \boldsymbol{\theta}) \sim (\mathbf{0}, \sigma^2 I) \\ &= B\boldsymbol{\theta} + \boldsymbol{U},\end{aligned}$$

- $\boldsymbol{Y} = (y_{11}, \cdots, y_{1T_1}, \cdots, y_{r1}, \cdots, y_{rT_r})' \in \mathbb{R}^{N\times 1}$
- $B_1 \in \mathbb{R}^{N\times r}$ is matrix of indicators for student $i = 1, ..., r$
- $B_2 \in \mathbb{R}^{N\times c}$ is matrix of indicators for the teachers $j = 1, ..., c$ matched to each student in each time period
- $B = [B_1, B_2]$: all the analysis is conditional on B
- $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$

Econometric Model

- Simplifications (for expositional purposes in the theory part of the talk):
  - Each period t, a student i is taught by a single teacher j.
  - In the examples: class size is constant $\kappa = r/c$ across teachers and time.
- Asymptotics:
  - T is fixed
  - $r, c \longrightarrow \infty$.

Next Steps in the Talk

1. Prior distribution with hyperparameter and posterior mean estimator

2. Hyperparameter selection based on minimization of unbiased risk estimate

3. Identification and optimality

4. (...)

Estimation: Overview

- Hierarchical Model and Empirical Bayes Method
  - Derive posterior mean estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ using an hierarchical prior
  $$p(\boldsymbol{\alpha}, \boldsymbol{\beta} | B, \boldsymbol{\lambda}).$$
  - Hyperparameter $\boldsymbol{\lambda}$ selection by minimization of an unbiased risk estimate (URE).
- Asymptotic Optimality
  - Frequentist risk (instead of integrated risk).
  - $\boldsymbol{\lambda}$ selection with URE minimization is robust to misspecification of prior distribution.

Estimation: Factorization of Prior

- Bayesian inference combines

$$\underbrace{p(B, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\lambda})}_{\text{prior}} \quad \text{and} \quad \underbrace{p(\boldsymbol{Y} | B, \boldsymbol{\alpha}, \boldsymbol{\beta})}_{\text{likelihood}}.$$

- From an economic perspective, the following factorization of the prior is natural and allows for sorting in the link formation:

$$p(B, \boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\lambda}) = p(\boldsymbol{\alpha}, \boldsymbol{\beta} | \boldsymbol{\lambda}) p(B | \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}).$$

- Because B is observed, posterior inference only requires

$$p(\boldsymbol{\alpha}, \boldsymbol{\beta} | B, \boldsymbol{\lambda}),$$

which we will use as our starting point.

Estimation: Prior

- Hyperparameter $\boldsymbol{\lambda} = (\mu, \lambda_\alpha, \lambda_\beta, \phi)$.
- Define

$$\Lambda = \begin{bmatrix} \lambda_\alpha \cdot I_r & 0 \\ 0 & \lambda_\beta \cdot I_r \end{bmatrix}, \quad D = \text{diag}(B'B), \quad \mathcal{A} = D^{-1/2}(B'B)D^{-1/2} - I.$$

- Prior Distribution:

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \Bigg| (B, \boldsymbol{\lambda}) \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{1}_r \mu \\ \mathbf{0}_c \end{bmatrix}, \sigma^2 \left[\Lambda^{1/2} \left(-\phi \mathcal{A} + I_{r+c} \right) \Lambda^{1/2} \right]^{-1} \right).$$

- No sorting for $\phi = 0$.

Estimation: Posterior Mean

- Shrinking the OLS estimator to common mean vector:

$$\hat{\boldsymbol{\theta}} = \mathcal{R} S_1(\boldsymbol{\lambda}) \hat{\boldsymbol{\theta}}^{LS} + \mathcal{R}\big(I - S_1(\boldsymbol{\lambda})\big)\boldsymbol{v}.$$

Estimation: Benchmark – Infeasible Oracle Shrinkage

- Consider Estimation of $\boldsymbol{\beta} \in \mathbb{R}^{c \times 1}$, e.g., teacher value added

- Quadratic Loss:

$$L(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \boldsymbol{\beta}) := \frac{1}{c} \sum_{j=1}^{c} \left(\hat{\beta}_j(\boldsymbol{\lambda}) - \beta_j\right)^2.$$

- Benchmark for Optimality (assumes known $\boldsymbol{\beta}$):

$$\hat{\boldsymbol{\beta}}^{OL}(\boldsymbol{\beta}) := \hat{\boldsymbol{\beta}}\left(\boldsymbol{\lambda}^{OL}(\boldsymbol{\beta})\right), \quad \boldsymbol{\lambda}^{OL}(\boldsymbol{\beta}) := \underset{\boldsymbol{\lambda} \in \Lambda}{\operatorname{argmin}}\, L(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \boldsymbol{\beta}).$$

Estimation: Feasible URE Shrinkage

- Frequentist Risk:

$$R(\boldsymbol{\lambda}) = \mathbb{E}_{B,\boldsymbol{\theta}}[L(\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}), \boldsymbol{\beta})].$$

- We derive an unbiased risk estimate (function of the data) such that:

$$\mathbb{E}_{B,\boldsymbol{\theta}}[URE(\boldsymbol{\lambda})] = R(\boldsymbol{\lambda}).$$

- Proposed Shrinkage Estimator:

$$\hat{\boldsymbol{\beta}}^{URE} := \hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}^{URE}), \quad \boldsymbol{\lambda}^{URE} := \operatorname*{argmin}_{\boldsymbol{\lambda} \in \Lambda} \mathsf{URE}(\boldsymbol{\lambda}).$$

Estimation: Example – Regression with ID Problem

$$y_i = \beta_i x_n + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n, \quad x_n \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

	Loss (λ)	Unbiased Risk Estimate (λ)
	$\frac{1}{n}\sum_{i=1}^n \left(\frac{x_n y_i}{x_n^2+\lambda} - \beta_i\right)^2$	$\frac{1}{(x_n^2+\lambda^2)^2}\left[\frac{\lambda^2}{x_n^2}\left(\frac{1}{n}\sum_{i=1}^n y_i^2 - \sigma^2\right) + x_n^2\sigma^2\right]$
$\lambda = 0$	$\frac{1}{x_n^2}\frac{1}{n}\sum_{i=1}^n u_i^2$	$\frac{1}{x_n^2}\sigma^2$
$\lambda = \infty$	$\frac{1}{n}\sum_{i=1}^n \beta_i^2$	$\frac{1}{n}\sum_{i=1}^n \beta_i^2 + \frac{2}{x_n}\frac{1}{n}\sum_{i=1}^n \beta_i u_i + \frac{1}{x_n^2}\frac{1}{n}\sum_{i=1}^2 (u_i^2 - \sigma^2)$

- Loss and URE at $\lambda = 0$ diverge.
- Need to control rate at which identification vanishes ($x_n \longrightarrow 0$), to ensure URE minimization is asymp. equivalent to loss minimization.

Identification: Some Intuition

Model:

$$y_{it} = \alpha_i + \beta_{j(i,t)} + u_{it}.$$

- Identification of β_j relies on students moving from teacher to teacher. Suppose student i is taught by teacher $j = t$ in period t:

$$y_{it} = \alpha_i + \beta_t + u_{it}, \quad t = 1, \ldots, T$$

$\Longrightarrow$ we learn that $y_{it} - \frac{1}{T}\Sigma_{j=1}^{T} y_{it} = \beta_t - \frac{1}{T}\Sigma_{j=1}^{T}\beta_t$ + noise.

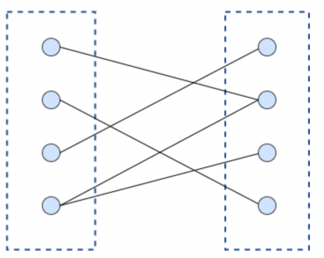
- We will assume that class size stays bounded which means that the c β_js cannot be consistently estimated. Under $r = \kappa c$ in the best case:

$$\frac{\#\text{ of equations}}{\#\text{ of parameters}} = \frac{rT}{r + c - 1} = \frac{\kappa T}{(\kappa + 1) - 1/c} \not\to \infty \quad \text{as} \quad c \to \infty.$$

Identification: Graph-theoretic Interpretation

- Bipartite Graph for θ

  - students ($\mathcal{S}$) and teachers ($\mathcal{T}$) on two sides
  - identification of θ requires a connected graph
  - connected graph: $\lambda_1(B'B) = 0$ and $\lambda_2(B'B) > 0$. i.e., the smallest eigenvalue is 0 and the second smallest eigenvalue is positive.
  - $\hat{\boldsymbol{\theta}}^{ls} = (B'B)^{-}B'\boldsymbol{Y}$



- Jochmans and Weidner (2019): $\lambda_2(B'B)$ measures global connectivity and determines properties of the OLS estimator, together with local measures.

Identification of $\beta_1, \ldots, \beta_c$: A Projected Graph

- FWL Theorem: OLS of $\boldsymbol{\beta}$ in $\boldsymbol{Y} = B_1\boldsymbol{\alpha} + B_2\boldsymbol{\beta} + \boldsymbol{U}$ is equivalent to OLS in:

$$\tilde{\boldsymbol{Y}} = [I - B_1(B_1'B_1)^{-1}B_1']\boldsymbol{Y} = [I - B_1(B_1'B_1)^{-1}B_1']B_2\boldsymbol{\beta} + \tilde{\boldsymbol{U}}.$$

- Define

$$B_{2,\perp} := [I - B_1(B_1'B_1)^{-1}B_1']B_2.$$

- Roughly:

$$\hat{\boldsymbol{\beta}}_{OLS} = \left(B_{2,\perp}'B_{2,\perp}\right)^{\dagger}B_{2,\perp}'\tilde{\boldsymbol{Y}}, \quad \text{eigv}_{(1)}(B_{2,\perp}'B_{2,\perp}) = 0.$$

- $B_{2,\perp}'B_{2,\perp}$ is adjacency matrix for projected graph that connects teachers through common students.
- Limited mobility: $\text{eigv}_{(2)}(B_{2,\perp}'B_{2,\perp})$ is close to zero $\Longrightarrow$ weak identification.

Identification of $\beta_1, \ldots, \beta_c$: Example

- Example: $c = 3$ teachers; class size κ; T time periods.
- $\nu \leq \kappa$ movers between periods $T-1$ and T; students move as follows:

$$j = 1 \mapsto j = 2, \quad j = 2 \mapsto j = 3, \quad j = 3 \mapsto j = 1.$$

- It can be shown that eigenvalues of $B'_{2,\perp} B_{2,\perp}$ are

$$\text{eigv}_{(1)} = 0, \quad \text{eigv}_{(2)} = 3\nu(1 - 1/T), \quad \text{eigv}_{(3)} = 3\nu(1 - 1/T).$$

  - If there are no movers ($\nu = 0$) $\text{eigv}_{(2)} = 0$ (no identification).
  - The more movers ν, the larger $\text{eigv}_{(2)}$.

- In the subsequent theory, we will impose conditions on $\text{eigv}_{(2)}(B'_{2,\perp} B_{2,\perp})$ to control strength of identification.

Main Theoretical Result: Optimality

Key regularity condition: There exists $j^* < \infty$ and $\delta > 0$ such that for $2 \leq j \leq j^*$, $\lim_{c \to \infty} c \cdot \text{eigv}_{(j)}(B_{2,\perp}' B_{2,\perp}) \to \infty$ and for $j > j^*$, $\text{eigv}_{(j)}(B_{2,\perp}' B_{2,\perp}) > \delta$.

Theorem (Asymptotic Optimality of URE Shrinkage)

Suppose (A1)–(A4) hold. Then for any $\epsilon > 0$,

$$\lim_{r,c \to \infty} \mathbb{P}_{B,\boldsymbol{\theta}} \left\{ L(\hat{\boldsymbol{\beta}}^{URE}, \boldsymbol{\beta}) \geq L(\hat{\boldsymbol{\beta}}^{OL}(\boldsymbol{\beta}), \boldsymbol{\beta}) + \epsilon \right\} = 0.$$

Implications: $\hat{\boldsymbol{\beta}}^{URE}$

- ... achieves (in probability) the same loss as $\hat{\boldsymbol{\beta}}^{OL}$;
- ... is asymptotically optimal among all feasible estimators within the class, e.g., EB-MLE, EB-MoM, OLS.

Simulation: Sorting and Connectivity

- Period $t = 0$: Draw (iid): $\alpha_i \sim \mathcal{N}(0, \sigma_a^2)$ and $\beta_j \sim N(0, \sigma_b^2)$.
- Period $t = 1$:
  (a) Allocate teachers and students to schools:
    - Sort teachers based on b_j draws. School 1 gets the worst teachers, ...
    - Re-assign a fraction of ρ teachers randomly across schools.
    - Sort students based on a_i draws. School 1 gets the worst students, ...
    - Re-assign a fraction of ρ students randomly across schools.

  (b) Match students to teachers:
    - Sort students within school based on a_i. Teacher 1 gets the worst students, ...
    - Re-assign a fraction of ρ students randomly across teachers.
- Period $t = 2$: a fraction ψ of randomly assigned teachers switches schools. Repeat student-to-teacher assignment (b).
- Generate outcomes $y_{it} = \alpha_i + \beta_{j(i,t)} + u_{it}$ conditional on graph $\mathcal{G}$.

ρ controls student-teacher sorting; ψ controls connectivity of graph.

Simulation: Data Generation & Estimators

- **Data generation:**
  - Parameters: 2000 students, 200 teachers across 20 schools. $\sigma_a^2 = 1$, $\sigma_b^2 = 1$.
  - Design 1 (uncorrelated effects): $\rho = 1$, $\psi = 0.2$.
  - Design 2 (correlated effects): $\rho = 0.5$, $\psi = 0.2$.
  - $N_{sim} = 500$ Monte Carlo repetitions.
- **Estimators:**

  OL $\boldsymbol{\lambda}$ selected using true loss (oracle);

  URE $\boldsymbol{\lambda}$ selected based on URE;

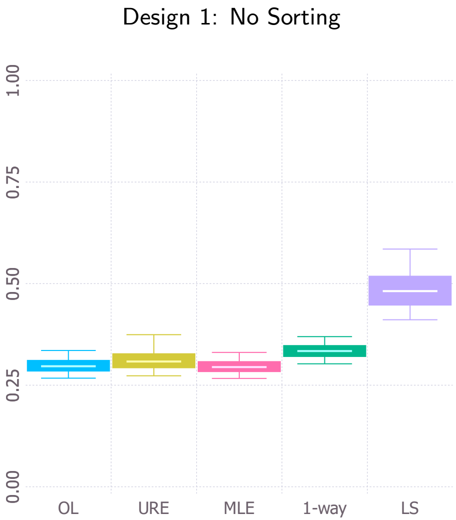
  MLE $\boldsymbol{\lambda}$ selected based on marginal likelihood;

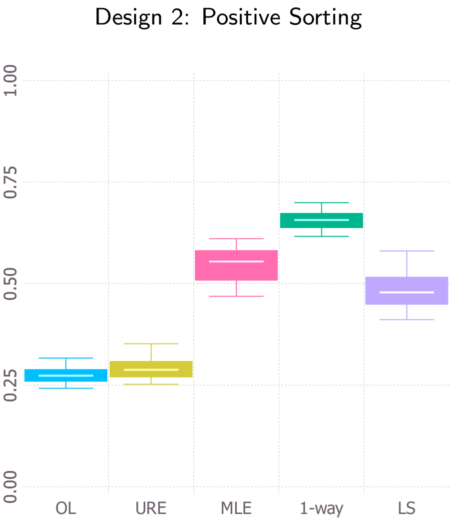
  1WAY one-way effects estimator;

  LS least squares.
- **Evaluation:**

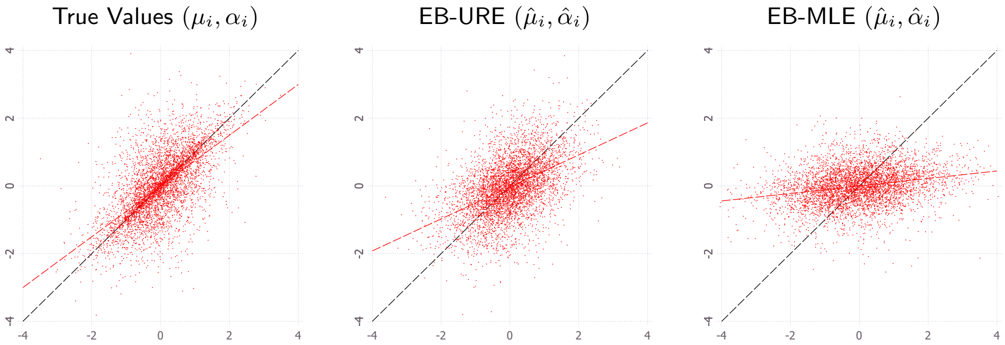
$$RMSE(\hat{\boldsymbol{\beta}}) = \sqrt{\frac{1}{N_{sim}} \sum_{s=1}^{N_{sim}} \frac{1}{c} \sum_{j=1}^{c} (\hat{\beta}_j^{(s)} - \beta_j)^2}$$

Simulation: URE λ Selection is Robust to Sorting



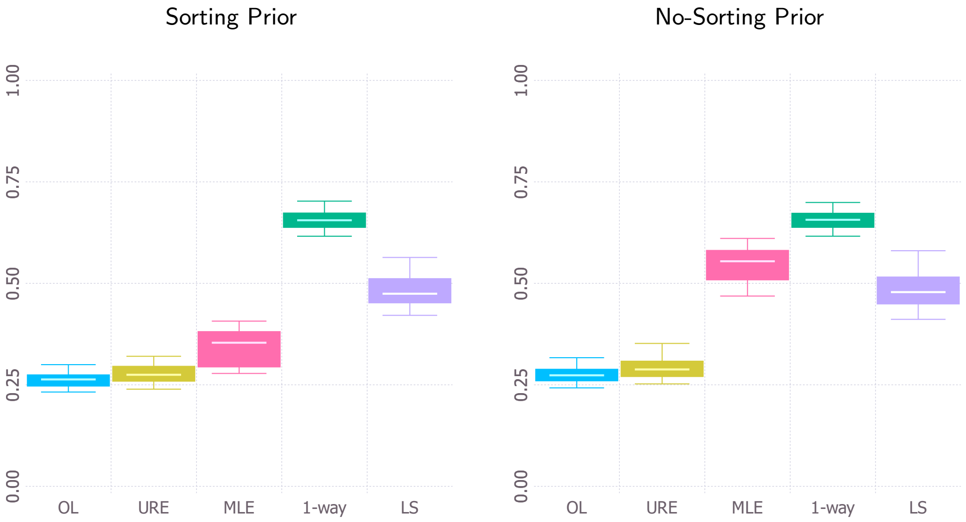


Simulation: URE-Based $(\hat{\alpha}, \hat{\beta})$ Estimates Capture Sorting

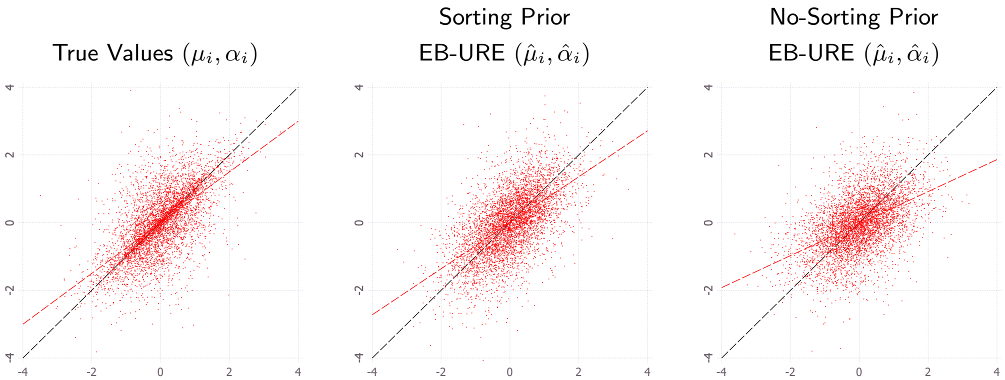


Notes: x-axis is μ_i and y-axis is α_i. Black lines are 45-degree lines and red lines are least squares regression lines. $\mu_i = 0.5(\beta_{j(i,1)} + \beta_{j(i,2)})$.

Simulation: Sorting Prior Improves MLE in Design 2



Simulation: $(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ Improve With Sorting Prior in Design 2



Notes: x-axis is μ_i and y-axis is α_i. Black lines are 45-degree lines and red lines are least squares regression lines. $\mu_i = 0.5(\beta_{j(i,1)} + \beta_{j(i,2)})$.

Empirical Application: In Progress

Matched student-teacher dataset from North Carolina Education Research Data Center (NCERDC)

- Sample: students from grades 3 to 5 for from 2018-2019.
- Outcome: math test score, standardized to have mean 0 and std 1 within (year,grade). Raw scores: mean 500, std 50.
- Student demographics: economically disadvantaged, english learner, sex, ethnicity.
- Class/teacher characteristics: not used.
- Connectivity: restrict to largest connected component of student-teacher graph:
  - 171,488 observations: $r = 132,037$ students, $c = 5,169$ teachers, and $s = 256$ schools.
  - 0.1 percentile of eig$(B_{2,\perp}' B_{2,\perp})$: 0.03 across all schools, 0.102 within schools.

$\Longrightarrow$ limited mobility across schools.

Empirical Application: Model

Quasi likelihood function:

$$y_{it} = 0.105 \cdot y_{it-1} + \alpha_i + \beta_{j(i,t)} + u_{it}, \quad u_{it} \sim \mathcal{N}(0, 0.121)$$

Prior Distribution:

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \Bigg| (B, \boldsymbol{\lambda}) \sim \mathcal{N}\left(\begin{bmatrix} X\boldsymbol{\gamma} \\ \mathbf{0}_c \end{bmatrix}, \ 0.121 \cdot \left[\Lambda^{1/2} \left(-\phi\mathcal{A} + I \right) \Lambda^{1/2} \right]^{-1} \right)$$

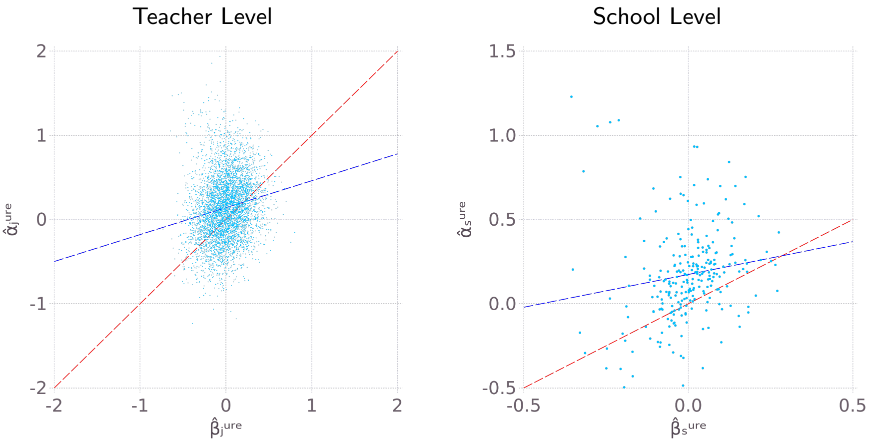
$$\Lambda = \begin{bmatrix} \lambda_\alpha \cdot I_r & 0 \\ 0 & \lambda_\beta \cdot I_r \end{bmatrix}, \quad D = \mathsf{diag}(B'B), \quad \mathcal{A} = D^{-1/2}(B'B)D^{-1/2} - I.$$

Empirical Application: Hyperparameter Estimates

- Shrinkage: $\hat{\lambda}_\alpha = 0.04$, $\hat{\lambda}_\beta = 1.1$.
- *A prior* sorting: $\hat{\phi} = 0.3$.
- Centering of student effects: $\hat{\gamma}$:

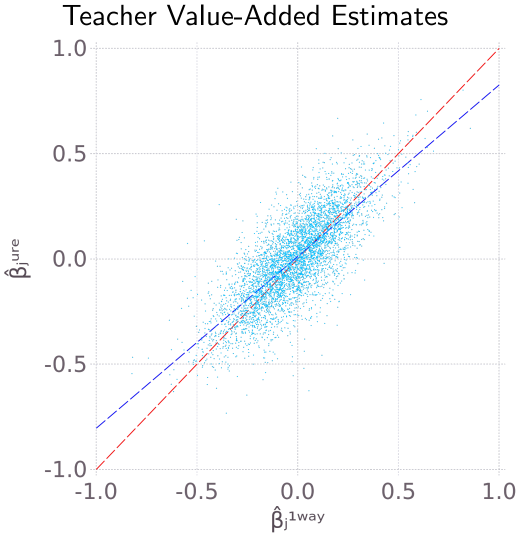
constant	-0.29
economically disadvantaged	-0.62
English learner	-1.85
female	0.07
asian	3.00
black	0.26
hispanic	1.12
white	0.67

Empirical Application: Positive Sorting



- $\hat{\alpha}_j^{ure}$: average of $\hat{\alpha}_i$ taught by teacher j.

Empirical Application: Two-way Versus One-way Effects



Teacher Value added Quintiles

One-way Effects	Two-way Effects 1st	2nd	3rd	4th	5th
1st	656	247	100	23	7
2nd	253	391	255	106	29
3rd	84	249	350	269	82
4th	31	111	242	386	264
5th	9	36	87	250	652

Concluding Remarks

- Two-way effects estimation with matched data.
- Scarce information manifests itself in weak connectivity of the student-teacher graph.
- A novel prior distribution that can capture sorting among students and teachers.
- Asymptotic optimality of URE-based hyperparameter selection for two-way shrinkage estimator.
- Application: teacher value-added estimation.
- Evidence for positive sorting at class and school level in NCERDC data set.