Empirical Bayes II: Empirical Applications (Part II)

ECON 2400: Applied Econometrics II

Juan C. Yamin

Brown University

Spring 2026

Outline

Outline

From theory to practice

Last session

- statistical decision theory and Bayes rules
- normal-normal empirical Bayes
- James-Stein shrinkage and the logic of borrowing strength
- a brief preview of nonparametric empirical Bayes

Today

- what EB looks like in real empirical applications
- why the first-step estimator matters
- why the relevant posterior object depends on the empirical objective

The common EB workflow

Three-step recipe

1. **Effect estimation:** estimate one parameter per unit, along with its standard error.
2. **Estimate a prior:** learn the cross-unit distribution from the whole ensemble.
3. **Posterior formation:** use the estimated prior to refine the estimate for each unit.

Key idea

- The estimate for unit j is interpreted using the information contained in the other units.
- The right posterior summary depends on the objective: estimation, classification, ranking, or reporting.

Outline

Application 1: School value-added

Consider a population of students indexed by i, each attending one of J schools.

Let $Y_i(j)$ denote student i's potential achievement if assigned to school j:

$$Y_i(j) = \beta_j + \varepsilon_i.$$

- β_j is the value-added of school j.
- ε_i captures student heterogeneity such as family background, prior preparation, and ability.
- Under this constant-effects model,

  $$\beta_j - \beta_k$$

  is the effect of moving any student from school k to school j.

Questions about schools

In this setting, at least three questions arise.

- A particular unit: what is the value-added of school j?
- The distribution: how much do schools differ? How wide is the distribution of β_j's?
- Decisions: which schools should be expanded, closed, or flagged for intervention?

Why EB is useful here

It helps with all three: learning the distribution, improving unit-level estimates, and supporting decisions.

VAM regression

Let D_{ij} indicate attendance at school j. Then observed achievement can be written as

$$Y_i = \sum_{j=1}^{J} \beta_j D_{ij} + \varepsilon_i.$$

Let observed covariates X_i, such as demographics and lagged scores, absorb part of student heterogeneity:

$$Y_i = \sum_{j=1}^{J} \beta_j D_{ij} + X_i'\gamma + u_i.$$

If selection on observables holds,

$$\mathbb{E}[u_i \mid D_{i1}, \ldots, D_{iJ}, X_i] = 0,$$

then OLS identifies the school effects in this value-added model.

What comes out of the VAM?

For each school j, VAM estimation gives

$$\{\hat{\beta}_j, s_j\}_{j=1}^{J},$$

that is,

- an estimated school effect $\hat{\beta}_j$,
- and its standard error s_j.

A useful approximation is

$$\hat{\beta}_j \mid \beta_j, s_j \sim \mathcal{N}(\beta_j, s_j^2).$$

Key feature

Different schools are estimated with different precision: some are based on much noisier evidence than others.

Why introduce a prior G?

Describe cross-school heterogeneity by a distribution

$$\beta_j \sim G.$$

What G captures

- how much does school value-added vary across schools?
- how dispersed are the latent β_j's after removing sampling noise?
- which schools are genuinely unusual relative to the rest?

Interpretation for this class

Think of G as an empirical description of heterogeneity across units. It gives a disciplined way to borrow strength from the ensemble.

Normal-normal empirical Bayes

Assume

$$\hat{\beta}_j \mid \beta_j, s_j \sim \mathcal{N}(\beta_j, s_j^2), \qquad \beta_j \mid s_j \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2).$$

Then the posterior mean is

$$\beta_j^* = \mathbb{E}[\beta_j \mid \hat{\beta}_j, s_j] = \frac{\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}\hat{\beta}_j + \frac{s_j^2}{\sigma_\beta^2 + s_j^2}\mu_\beta.$$

Interpretation

- Noisier schools shrink more.
- Precise schools shrink less.
- The posterior mean combines school-specific and population-level information.

Angrist et al. (2017): motivation

School value-added is used for high-stakes decisions: closures, expansions, interventions, and public reporting.

In practice, three problems arise:

- **Selection bias:** schools may look good because they enroll stronger students.
- **Sampling noise:** some schools are measured much less precisely than others.
- **Limited experimental variation:** lottery evidence is more credible, but only available for some schools and often much noisier.

Core question: Can lottery evidence help us assess conventional VAMs and construct better school-level estimates?

Boston setting and data

Boston students choose among several school sectors:

- **Traditional BPS schools:** standard district-run public schools.
- **Pilot schools:** district schools with more autonomy
- **Charter schools:** independently operated public schools

Admissions generate useful quasi-random variation:

- traditional and pilot schools are assigned through a centralized deferred-acceptance match,
- ties are broken using random lottery numbers,
- oversubscribed charter schools also use admissions lotteries.

Empirically, the paper studies roughly 28,000 sixth-grade students in 51 Boston schools over the 2006–2007 to 2013–2014 school years.

Step 1: Conventional OLS value-added

For student i, estimate school effects from

$$Y_i = \sum_{j=1}^{J} \alpha_j D_{ij} + X_i'\gamma + u_i,$$

where:

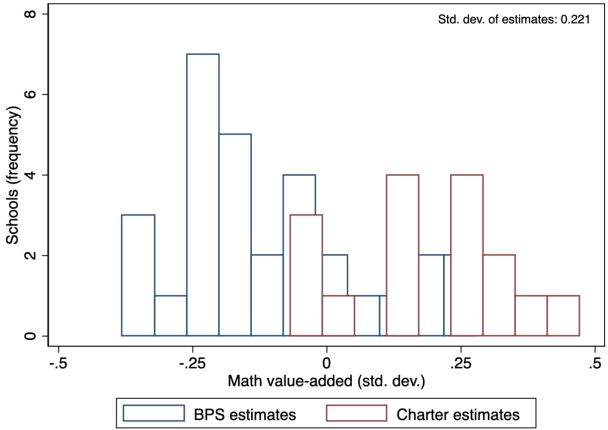
- Y_i is sixth-grade test achievement,
- D_{ij} indicates attendance at school j,
- X_i includes lagged scores and demographics.

The resulting OLS estimate $\hat{\alpha}_j$ is:

- relatively precise,
- but potentially biased if school attendance is still correlated with unobservables.

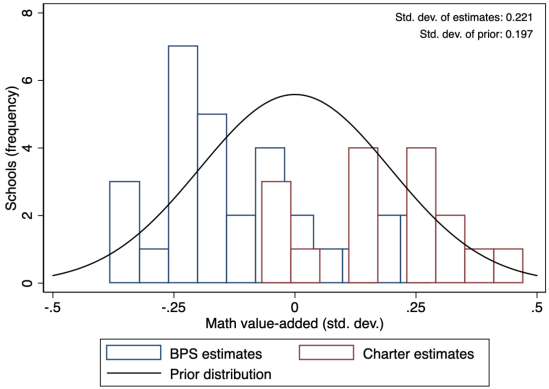
Output: one conventional VAM estimate $\hat{\alpha}_j$ for each school.

Boston example: raw lagged-score VAM estimates



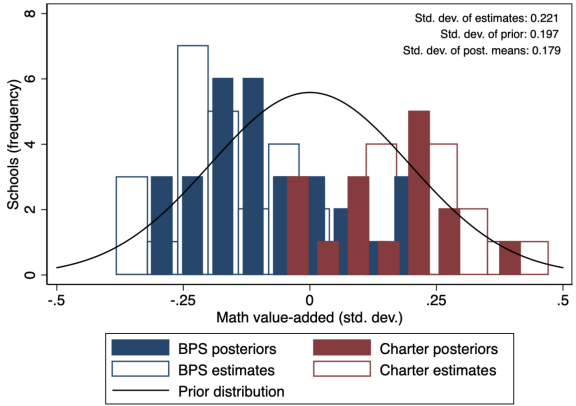
- Raw VAM estimates show substantial dispersion across schools.
- But this observed spread mixes true heterogeneity and sampling noise.

Prior Distribution Pooling Sectors



- Under a pooled prior, noisy school estimates shrink toward a common mean.

Posterior Means Pooling Sectors



- The posterior means are even tighter than the prior because they are shrinkage estimates, not the latent school effects themselves.

Incorporating covariates into the prior

A common prior can be too crude. Instead, the prior mean depends on school characteristics:

$$\beta_j \mid s_j, C_j \sim \mathcal{N}(C_j'\mu, \sigma_r^2).$$

C_j may include indicators for charter and pilot sectors: $C_j = (1, \text{Charter}_j)$.

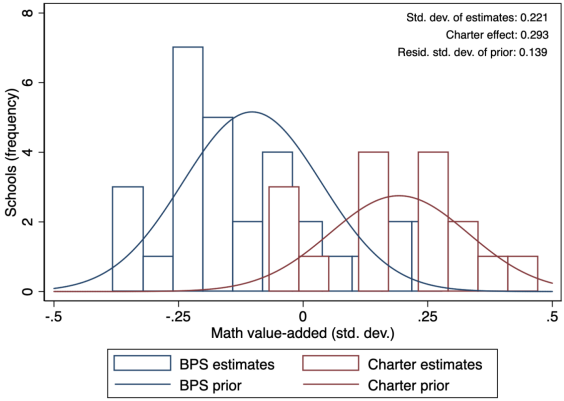
Implementation

1. Regress $\hat{\beta}_j$ on C_j to estimate the prior mean function $C_j'\hat{\mu}$.
2. Form residuals $\hat{r}_j = \hat{\beta}_j - C_j'\hat{\mu}$.
3. Estimate residual heterogeneity by: $\hat{\sigma}_r^2 \approx \text{Var}(\hat{r}_j) - \mathbb{E}[s_j^2]$
4. Shrink school j toward its covariate-specific mean $C_j'\hat{\mu}$

What changes?

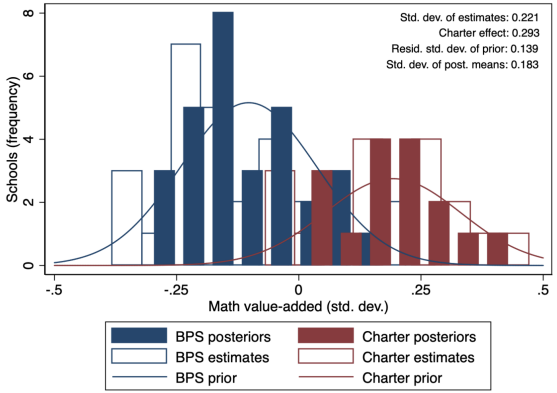
- Schools now learn more from other schools with similar observed characteristics.
- This avoids forcing BPS and charter schools to be shrunk toward the same benchmark when their average value-added may differ.

New priors, new posteriors



- Allowing a charter indicator shifts the prior mean to the right for charter schools.
- After removing this sector mean difference, the residual prior standard deviation falls to 0.139.

Posteriors Shrinking Toward Sector Means



- Each school is pulled toward its *sector-specific* prior mean.

From denoising to debiasing

So far, empirical Bayes helped us with **sampling noise** in school value-added estimates.

Angrist et al. (2017) add a new problem:

- conventional OLS VAM estimates are relatively precise,
- but they may be **biased** if selection on observables fails,
- while lottery-based estimates are more credible causally, but much noisier.

New question

How should we estimate school quality when one signal is precise but potentially biased, and another is credible but noisy?

A stylized hybrid EB problem

To build intuition, suppose we have two estimates of the same school effect β_j:

$$\hat{\beta}_j^{OLS} \mid \beta_j, b_j \sim \mathcal{N}(\beta_j + b_j,\ s_{j,OLS}^2),$$

$$\hat{\beta}_j^{L} \mid \beta_j, b_j \sim \mathcal{N}(\beta_j,\ s_{j,L}^2).$$

- β_j: causal school value-added,
- b_j: bias in the observational VAM,
- $\hat{\beta}_j^{OLS}$: relatively precise, but potentially biased,
- $\hat{\beta}_j^{L}$: lottery-based estimate, more credible but much noisier.

Empirical Bayes treats (β_j, b_j) as draws from a joint population distribution and uses the ensemble $\{(\hat{\beta}_j^{OLS}, \hat{\beta}_j^{L})\}_{j=1}^{J}$ to learn that distribution.

This is a simplified benchmark. In the actual paper, lottery evidence is more limited, so the implementation is more complicated.

What hybrid EB estimates

The hybrid posterior mean is

$$\hat{\beta}_j^H = \mathbb{E}[\beta_j \mid \hat{\beta}_j^{OLS}, \hat{\beta}_j^L, s_{j,OLS}, s_{j,L}].$$

It combines three sources of information:

- the noisy but credible lottery estimate $\hat{\beta}_j^L$,
- the precise but potentially biased OLS estimate $\hat{\beta}_j^{OLS}$,
- population-level information on how school quality and bias vary across schools.

Key idea

Do not throw away OLS just because it may be biased. If OLS is still informative about β_j, EB can use it while adjusting for bias.

A transparent special case

In a simplified benchmark, the hybrid posterior mean can be written as

$$\hat{\beta}_j^H = w_L \hat{\beta}_j^L + w_O(\hat{\beta}_j^{OLS} - \hat{b}) + (1 - w_L - w_O)\hat{\mu}_\beta,$$

where:

- $\hat{b}$ is the estimated average bias in the OLS signal,
- $\hat{\mu}_\beta$ is the estimated population mean of true school value-added,
- w_L and w_O are weights chosen to minimize MSE.

Interpretation

Hybrid EB is a weighted average of:

- a credible but noisy estimate,
- a precise but bias-adjusted estimate,
- and the prior mean.

What hybrid EB buys you

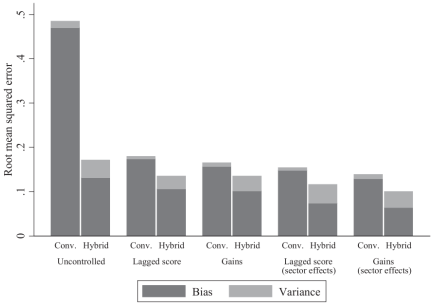


FIGURE VI

Root Mean Squared Error for Value-Added Posterior Predictions

This figure plots root mean squared error (RMSE) for posterior predictions of sixth-grade math value-added. Conventional predictions are posterior means constructed from OLS value-added estimates. Hybrid predictions are posterior modes constructed from OLS and lottery estimates. The total height of each bar indicates RMSE. Dark bars display shares of mean squared error due to bias, and light bars display shares due to variance. RMSE is calculated from 500 simulated samples drawn from the data generating processes implied by the estimates in Table VI. The random coefficients model is reestimated in each simulated sample.

How to read the figure

- Each specification compares **conventional** and **hybrid** posterior predictions.
- Dark shading is the share of MSE due to **bias**.
- Light shading is the share of MSE due to **variance**.

Main message

- Hybrid EB lowers RMSE relative to conventional VAM estimates.
- It mainly reduces **bias**, though at the cost of somewhat more **variance**.
- In Boston, that tradeoff is favorable and also improves policy targeting.

Decision-making for schools

Posterior means are optimal under squared-error loss.

But suppose the policy problem is **classification**:

> which schools should be flagged as sufficiently low value-added to warrant intervention?

Then the relevant posterior object is not necessarily the posterior mean.

Main point

We need to write down the loss function explicitly.

School classification under an explicit loss function

Let $\delta_j \in \{0, 1\}$ indicate whether school j is selected for intervention, and let c be the cutoff for low value-added.

Use the loss function

$$L(\beta_j, \delta_j) = \delta_j \, \mathbb{1}\{\beta_j > c\} + (1 - \delta_j) \, \mathbb{1}\{\beta_j \leq c\} \, \kappa.$$

Interpretation

- Intervene when $\beta_j > c \Rightarrow$ false positive, cost 1.
- Fail to intervene when $\beta_j \leq c \Rightarrow$ false negative, cost κ.

Optimal rule

$$\delta_j^* = \mathbb{1}\left\{ \mathbb{P}(\beta_j < c \mid \hat{\beta}_j, s_j) \geq \frac{1}{1 + \kappa} \right\}.$$

Why classification is different

For estimation under squared-error loss, the posterior mean is the natural summary.

For intervention, the question is different:

How likely is it that this school is truly below the policy cutoff c?

- This is a **tail-probability** problem, not a center-of-the-posterior problem.
- Two schools can have the same posterior mean but very different posterior risk of being below c.
- The parameter κ tells us how much posterior tail risk is enough to justify intervention.

Big takeaway

For MSE, summarize the posterior by its mean.
For classification, summarize the posterior by the mass in the relevant tail, or equivalently by a posterior quantile.

Outline

Application 2: Discrimination by large U.S. employers

Question: do major employers treat applicants differently by race at the very first stage of hiring?

- It studies a high-stakes outcome: whether employers contact an applicant after receiving a resume.
- It focuses on specific large firms, not just the labor market on average.
- It asks whether discrimination is diffuse across many firms or concentrated in a smaller set of employers.

Main substantive finding from the experiment:

- Distinctively Black names reduce employer contact by about 2.1 percentage points on average.
- But the average hides substantial heterogeneity across firms.

Why this is a natural EB setting: there are many firm-specific parameters, each measured with noise, and we care about both

- the distribution of discrimination across firms, and
- the conduct of particular firms.

Experimental design

The paper runs a large-scale resume correspondence experiment aimed at learning about *specific firms*, not just the market average.

Basic setup:

- 108 large employers
- up to 125 entry-level jobs per firm

What do they send?

- Fictitious resumes tailored to real entry-level vacancies
- Applications sent in pairs over time to the same job

Outcome: Whether the employer attempts to contact the applicant within 30 days

How does the randomization work?

Fix one job posting at firm f.

The researchers send ≈ 8 fictitious applications to that same vacancy, typically organized as four pairs.

Within each pair:

- one application is randomly assigned a distinctively White name
- the other is randomly assigned a distinctively Black name

So for a given job, the design aims to create a balanced race comparison:

4 White-signaling applications vs. 4 Black-signaling applications.

Each fictitious applicant is also randomly assigned other resume details (first and last name, educ., etc.)

Key idea:

- race is randomized *within the job*, subject to balance
- the job-level callback gap compares how the *same vacancy* responds to White- vs. Black applicants

Job-level estimates

Let $Y_{ijf}(r) \in \{0,1\}$ indicate whether applicant i would receive a callback from job j at firm f if assigned perceived race $r \in \{b, w\}$.

The job-level racial callback gap is

$$\Delta_{jf} \equiv \mathbb{E}[Y_{ijf}(w) - Y_{ijf}(b)] .$$

We observe $Y_{ijf} = Y_{ijf}(R_{ijf})$, $R_{ijf} \in \{b, w\}$, where perceived race is randomly assigned within the job.

Since each job receives about 4 White-signaling and 4 Black-signaling applications, a natural estimator is the within-job difference in callback rates:

$$\hat{\Delta}_{jf} = \frac{1}{4} \sum_{i=1}^{8} \mathbf{1}\{R_{ijf} = w\} Y_{ijf} - \frac{1}{4} \sum_{i=1}^{8} \mathbf{1}\{R_{ijf} = b\} Y_{ijf}.$$

By random assignment,

$$\mathbb{E}[\hat{\Delta}_{jf}] = \Delta_{jf}.$$

So $\hat{\Delta}_{jf}$ compares how the same vacancy responds to White- and Black-signaling applicants.

From jobs to firms

A single job-level estimate is still noisy. The paper therefore moves from job-level discrimination to firm-level discrimination.

Let the average racial callback gap across jobs within firm f be:

$$\Delta_f := \mathbb{E}_f[\Delta_{jf}]$$

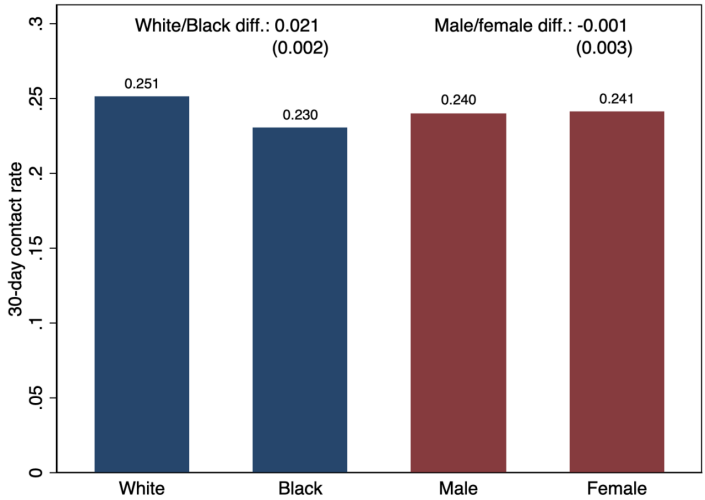
The observed firm-level estimate is

$$\hat{\Delta}_f = \frac{1}{J_f} \sum_{j=1}^{J_f} \hat{\Delta}_{jf},$$

where J_f is the number of sampled jobs for firm f. Under random assignment of names and random sampling of jobs within firms, $\mathbb{E}[\hat{\Delta}_f] = \Delta_f$.

- $\hat{\Delta}_{jf}$ tells us whether one vacancy appears discriminatory.
- $\hat{\Delta}_f$ asks whether the employer systematically favors White- over Black-signaling applicants across its hiring process.

Average Contact Gaps by Race and Gender



The distribution of discrimination

How is discrimination distributed across employers?

Let G denote the distribution of firm-level contact gaps:

$$\Delta_f \sim G(\Delta), \qquad f = 1, \ldots, F.$$

This object is substantively interesting because it answers questions like:

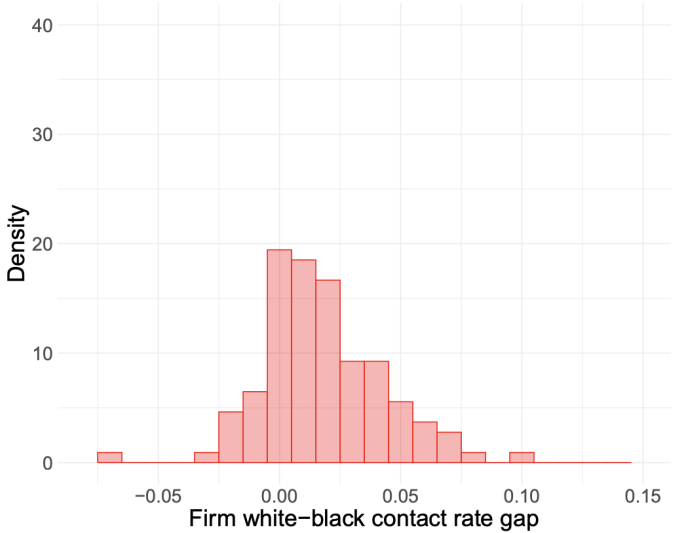
- Is the average White–Black callback gap driven by many mildly discriminatory firms?
- Or by a smaller set of severe discriminators?

But we do not observe Δ_f directly. We observe noisy estimates:

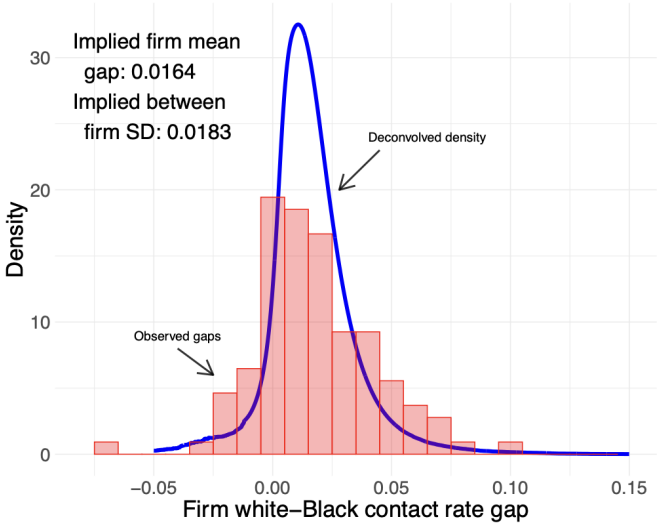
$$\hat{\Delta}_f \mid \Delta_f, s_f \sim N(\Delta_f, s_f^2).$$

So the empirical Bayes problem is: use the noisy firm-level estimates $\{(\hat{\Delta}_f, s_f)\}_{f=1}^{F}$ to learn the underlying distribution G and then compute posterior distribution!

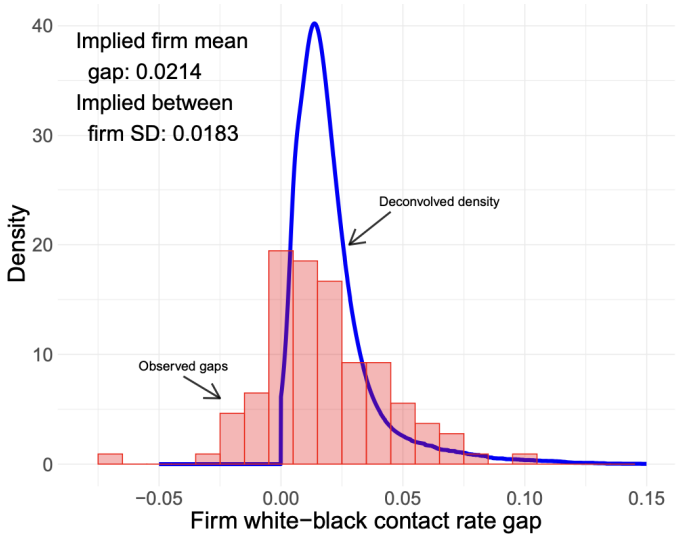
Histogram of Race Contact Gap Estimates



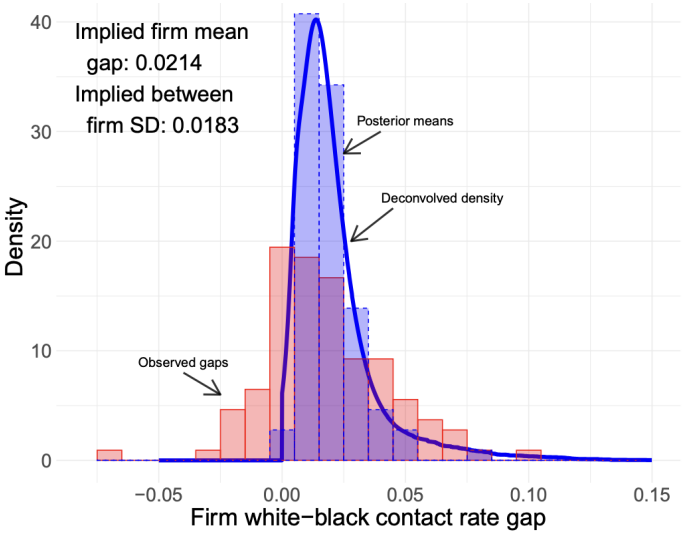
Histogram of Race Contact Gap Estimates



Histogram of Race Contact Gap Estimates



Histogram of Posterior Mean



Different goals need different posterior summaries

If the goal is estimation:

- We want a good estimate of the firm's discrimination level.
- Then the posterior mean is natural.

If the goal is classification:

- We want to know whether a firm is likely discriminating at all.
- Then the key object is not the center of the posterior, but the posterior mass above zero.

So the question changes from

$$\text{How large is } \Delta_f \text{ on average?}$$

to

$$\text{How likely is it that } \Delta_f > 0 \text{ ?}$$

That is why posterior means are useful for estimation, but posterior probabilities become more useful for flagging firms.

Why q-values?

Suppose we test, for each firm,

$$H_0 : \Delta_f = 0 \qquad \text{vs.} \qquad H_A : \Delta_f > 0.$$

With many firms, the problem is no longer just whether one firm has a small p-value.

The real question is:

> *If we flag a set of firms as likely discriminators, how many mistakes should we expect inside that set?*

This is what the false discovery rate controls.

- A small p-value means strong evidence against the null for one firm.
- A small q-value means strong evidence *and* a low expected mistake rate among all flagged firms.

So q-values are useful when the goal is not just testing one firm, but identifying a set of firms while limiting false accusations.

What do they find?

Using large-scale multiple-testing methods, the paper finds that:

- 23 firms can be classified as discriminating against Black applicants while controlling the false discovery rate at 5%
- those firms account for nearly 40% of lost contacts to Black applicants in the experiment

Interpretation:

- empirical Bayes is useful not only for estimating firm effects
- it is also useful for making selective decisions in a way that controls mistakes

Outline

Same EB pipeline, different empirical goals

	Schools	**Employers**
First-step object	OLS and lottery VA estimates	Experimental contact gaps
Prior	Parametric or covariate-adjusted	Flexible nonparametric
Posterior use	Shrinkage, bias correction, classification	Shrinkage, classification, reporting
Decision problem	Which schools to flag or intervene on	Which firms to investigate or how to grade them

Final takeaways

What we learned today

1. Empirical Bayes starts with noisy unit-level estimates and their standard errors.
2. The prior is learned from the ensemble of units.
3. Posterior means are only one possible use of the posterior.
4. When the goal is classification or reporting, the loss function should be made explicit.
5. In applied work, EB is useful for smoothing, bias correction, targeting, testing, and public reporting.

The throughline

Same EB skeleton across applications: noisy unit-level estimates $\rightarrow$ ensemble-based prior $\rightarrow$ posterior summary chosen to match the objective.