Empirical Bayes and Large-Scale Inference

Mixtape Session

Prof. Christopher Walters

Empirical Bayes and Large-Scale Inference

- Economists are increasingly drilling down to study heterogeneity in fine-grained, unit-specific parameters
  - Returns to a year of education $\implies$ Returns to college selectivity $\implies$ Returns to specific colleges (Card, 1999; Dale and Krueger, 2002, 2014; Mountjoy and Hickman, 2021)
  - Industry wage premia $\implies$ Firm-specific wage premia (Krueger and Summers, 1988; Abowd et al., 1999; Card et al., 2018)
  - Effects of neighborhood characteristics $\implies$ Effects of specific neighborhoods (Kling et al., 2007; Chetty and Hendren, 2018; Chetty et al., 2018)
  - "Value-added" of individual teachers, schools, doctors, medical centers, managers, police officers, judges (Chetty et al., 2014; Angrist et al., 2017; Chan et al., 2022; Einav et al., 2022; Fenizia, 2022; Goncalves and Mello, 2023; Frandsen et al., 2023)

Empirical Bayes and Large-Scale Inference

- In settings with many unit-specific parameters, **empirical Bayes (EB)** methods are useful for several purposes
  - Learning about the *distribution* of parameters across units
  - Improving estimates for individual units ("borrowing strength")
  - Making decisions (Policy: what to do? Scientific: what to report?)
- The goals of this session are to familiarize students with EB methods and provide tools to apply them in practice

Course Outline

- Part 1: Empirical Bayes basics
  - Conceptual framework and empirical Bayes recipe
  - Linear shrinkage; James/Stein theorem
  - Uses of EB: combining estimators, EB and regression, EB decision rules, individualized treatment recommendations
  - Comparisons to other methods: Machine learning, full B vs. EB
- Part 2: Empirical Bayes extensions
  - Flexible variance estimation
  - Precision-dependence
  - Non-parametric deconvolution and shrinkage
- Part 3: Large-scale inference
  - Empirical Bayes approaches to False Discovery Rate control
- Each part will include a coding lab to illustrate the methods, along with live-coding of solutions

Conceptual Framework

- We'll start with a basic conceptual framework, illustrated through a running example
- Let $j \in \{1...J\}$ index groups (e.g. schools), and let $i \in \{1...N\}$ index individuals within groups (e.g. students)
- We observe outcomes Y_{ij} for each individual i in each group j
- θ_j is an unknown parameter for group j that determines the distribution of Y_{ij} (e.g. the effect of school j)
- For example, suppose test scores are normally distributed and homoskedastic, with a school-specific mean:

$$Y_{ij} = \theta_j + \epsilon_{ij}, \ \epsilon_{ij}|\theta_j \sim N(0, \sigma_\epsilon^2)$$

Estimating θ_j

- We start by using the data for group j to form an estimate $\hat{\theta}_j$ of θ_j
- In our school effects example, the natural estimator is the sample mean for each school:

$$\hat{\theta}_j = \frac{1}{N}\sum_{i=1}^{N} Y_{ij}$$

- The $\hat{\theta}_j$'s are unbiased, noisy estimates of the unknown θ_j's:

$$\hat{\theta}_j|\theta_j \sim N\left(\theta_j, s_j^2\right)$$

- (Squared) standard errors $s_j^2 \equiv Var\left(\hat{\theta}_j|\theta_j\right)$ quantify the noise in $\hat{\theta}_j$
  - With equal group sizes and homoskedastic errors, $s_j^2 = \sigma_\epsilon^2/N \ \forall j$

Introducing G

- Next, we posit a distribution of the parameters θ_j across groups:

$$\theta_j \sim G,\ j = 1, ..., J$$

- The **mixing distribution** G is a key object in the EB framework
- G is an objective feature of the world, not a subjective prior
- G answers questions about variation in parameters
  - How much does school quality vary? $\sigma^2_\theta = \int (\theta - \mu_\theta)^2 dG(\theta)$
  - What's the difference between 75th and 25th percentile schools? $G^{-1}(0.75) - G^{-1}(0.25)$
- EB **deconvolution**: Use noisy estimates $\hat{\theta}_j$ along with standard errors s_j to compute an estimate $\hat{G}$ of G

The Philosophy of G

- What does it mean to treat the θ_j parameters as random draws from a distribution G?
  - "Fixed effects" perspective: There are J units, with fixed but unknown parameters $\{\theta_j\}_{j=1}^{J}$
  - One (unsatisfying) answer: observed units are sampled from some larger superpopulation
- "Random effects" perspective can be motivated by analyst's objectives
  - Even with finite population of units, we can ask how the θ_j's are distributed in this population
  - If our loss function cares about average performance across units, it's valuable to incorporate distributional information into estimates for individuals
  - Think of continuous/*iid* models for G as parsimonious approximation
  - Random vs. fixed effects is *not* about correlation of θ_j's with observables (c.f. "random effects" vs. "correlated random effects")

Normal/Normal Model

- In school effects example, suppose G is a normal distribution and independent of s_j:
- Then we have the hierarchical model

$$\hat{\theta}_j|\theta_j, s_j \sim N(\theta_j, s_j^2)$$

$$\theta_j|s_j \sim N(\mu_\theta, \sigma_\theta^2)$$

- **Hyperparameters** μ_θ and σ_θ^2 summarize the distribution of lower-level parameters
- With this model for G, deconvolution just requires estimating these two hyperparameters

Estimating Hyperparameters

- Simple estimators for hyperparameters:

$$\hat{\mu}_\theta = \frac{1}{J}\sum_{j=1}^{J} \hat{\theta}_j,$$

$$\hat{\sigma}^2_\theta = \frac{1}{J}\sum_{j=1}^{J} \left[(\hat{\theta}_j - \hat{\mu}_\theta)^2 - s_j^2\right]$$

- Subtracting the average squared standard error s_j^2 is a **bias-correction** accounting for excess variance in the $\hat{\theta}_j$'s due to sampling error
  - $\hat{\sigma}^2_\theta > 0 \implies$ **overdispersion** beyond what we'd expect from noise
- Other options: precision-weighting; maximum likelihood estimation (MLE)

Linear Shrinkage

- In normal/normal model, posterior mean for θ_j given $\hat{\theta}_j$ is:

$$\theta_j^* \equiv E[\theta_j|\hat{\theta}_j] = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\sigma_\theta^2 + s_j^2}\right)\mu_\theta$$

- Posterior mean **shrinks** noisy estimate $\hat{\theta}_j$ toward prior mean based on signal-to-noise ratio
- Linear shrinkage formula is motivated by normality, but has good properties regardless of the form of G
  - Coincides with fitted value from OLS regression of θ_j on $\hat{\theta}_j$ $\implies$ inherits usual OLS "best linear predictor" interpretation

EB Posterior Means

- Putting the "E" in "EB" – Empirical Bayes posterior mean $\hat{\theta}_j^*$ plugs in estimated hyperparameters:

$$\hat{\theta}_j^* = \left(\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2 + s_j^2} \right) \hat{\theta}_j + \left(\frac{s_j^2}{\hat{\sigma}_\theta^2 + s_j^2} \right) \hat{\mu}_\theta$$

- EB posterior shrinks estimate for school j using hyperparameters estimated with the larger pool of schools
- Reflects general EB approach: Use deconvolution estimate $\hat{G}$ as prior when forming posteriors for individual units
  - "Borrowing strength from the ensemble" (Efron and Morris, 1973; Morris, 1983)
  - "Indirect evidence" (Efron, 2010)
  - "Learning from the experience of others" (Efron, 2012)

Summary: A Three-step EB Recipe

1. **Estimation:** Estimate parameter and standard error for each unit $\implies \{\hat{\theta}_j, s_j\}_{j=1}^J$
2. **Deconvolution:** Use $\{\hat{\theta}_j, s_j\}_{j=1}^J$ to estimate mixing distribution $\implies \hat{G}$
3. **Shrinkage:** Treating $\hat{G}$ as prior, update with $(\hat{\theta}_j, s_j)$ to form posterior $\implies \{\hat{\theta}_j^*\}_{j=1}^J$

When to Shrink?

- Should we prefer the shrunk posterior mean to the unbiased estimate $\hat{\theta}_j$? It depends on our goals
- Continue with the normal/normal model:

$$\hat{\theta}_j|\theta_j, s_j \sim N(\theta_j, s_j^2), \quad \theta_j|s_j \sim N(\mu_\theta, \sigma_\theta^2)$$

- Conditional on the latent parameter for school j, mean squared error (MSE) for the two estimators is:

$$E\left[(\hat{\theta}_j - \theta_j)^2|\theta_j, s_j\right] = s_j^2$$

$$E\left[(\hat{\theta}_j - \theta_j)^2|\theta_j, s_j\right] = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s_j^2}\right)^2 s_j^2 + \left(\frac{s_j^2}{\sigma_\theta^2+s_j^2}\right)^2 (\theta_j - \mu_\theta)^2$$

- If we're only interested in one school (e.g. θ_1), not clear which is better
- Shrinkage reduces variance, but may introduce substantial bias if the school is very different from average

Shrinkage Reduces Aggregate MSE

- Now suppose we're equally interested in all schools
- In this case the relevant notion of MSE integrates over G:

$$E\left[(\hat{\theta}_j - \theta_j)^2 | s_j\right] = \int E\left[(\hat{\theta}_j - \theta)^2 | \theta_j = \theta, s_j\right] dG(\theta) = s_j^2$$

$$E\left[(\theta_j^* - \theta_j)^2 | s_j\right] = \int E\left[(\theta_j^* - \theta)^2 | \theta_j = \theta, s_j\right] dG(\theta) = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right) s_j^2$$

- Linear shrinkage estimate is superior if we want an estimator that performs well *on average across schools*
- Related to classic **James/Stein** (1961) result: sample mean is inadmissible under squared loss and dominated by shrinkage-based estimators when estimating at least three parameters

The James/Stein Theorem

- Dispense with random effects framework for a moment, and suppose we're interested in $J \geq 3$ parameters $(\theta_1, ..., \theta_J)'$
- Let $\hat{\theta}_j | \theta_j \sim N(\theta_j, 1)$, independent across j
- We're interested in choosing an estimator with low total MSE across all J parameters
- MSE of unbiased estimates:

$$MSE_{\hat{\theta}} = \sum_{j=1}^{J} E\left[(\hat{\theta}_j - \theta_j)^2 | \theta_j\right] = J$$

The James/Stein Theorem

- Consider an alternative shrinkage estimator of the form:

$$\hat{\theta}_j^{JS} = \left(1 - \frac{(J-2)}{\sum_{k=1}^{J}(\hat{\theta}_k - m)^2}\right)\hat{\theta}_j + \left(\frac{(J-2)}{\sum_{k=1}^{J}(\hat{\theta}_k - m)^2}\right)m$$

- This estimator shrinks $\hat{\theta}_j$ toward a constant m, using sum of squares to choose the shrinkage factor
- $\hat{\theta}_j^{JS}$ dominates the unbiased estimator $\hat{\theta}_j$ on MSE grounds, as shown by the **James/Stein (1961) Theorem**:

$$MSE_{\hat{\theta}^{JS}} = E\left[\sum_j \left(\hat{\theta}_j^{JS} - \theta_j\right)^2 |\theta_j\right]$$

$$\leq J - \frac{(J-2)^2}{J - 2 + \sum_{j=1}^{J}(\theta_j - m)^2}$$

$$< J$$

$$= MSE_{\hat{\theta}}.$$

James/Stein Discussion

- JS shrinkage emerges naturally as EB procedure based on normal G with known mean m and unknown variance
  - If $\theta_j \sim N(m, \sigma_\theta^2)$, then $(J-2)/[\sum_j(\hat{\theta}_j - m)^2]$ is an unbiased estimate of shrinkage factor $1/(1+\sigma_\theta^2)$
- But MSE improvement does not require random effects/Bayesian framework – holds for any configuration of θ_j's and any m
- Result generalizes to using estimated standard error $\neq 1$ and estimated prior mean
- We can often think of mixing distribution G as a device to motivate procedures with desireable frequentist properties
- Note that JS result requires us to care equally about each of the $J \geq 3$ parameters – no guarantee of improvement for any particular θ_j
  - Should we jointly shrink estimates of the age of the universe, elasticity of labor supply, and efficacy of Covid vaccine?

Unbiased Estimates vs. Shrunk Posteriors vs. True Effects

- It's important to keep in mind the distinction between unbiased estimates $\hat{\theta}_j$, linear shrinkage estimates $\hat{\theta}_j^*$, and true latent parameters θ_j
- If we're interested in understanding variation in the true θ_j's, the variance of unbiased estimates $\hat{\theta}_j$ is *too big:*

$$\begin{aligned} Var(\hat{\theta}_j) &= \sigma_\theta^2 + s_j^2 \\ &> \sigma_\theta^2. \end{aligned}$$

- On the other hand, the variance of shrunk posteriors θ_j^* is *too small:*

$$\begin{aligned} Var\left(\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\sigma_\theta^2+s_j^2}\right)\mu_\theta\right) &= \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s_j^2}\right)\sigma_\theta^2 \\ &< \sigma_\theta^2. \end{aligned}$$

- Deconvolution estimate $\hat{\sigma}_\theta^2$ is the "goldilocks variance" that gets it just right – consistent estimate of σ_θ^2
- More generally, to summarize features of the distribution of θ_j's, we should use an estimate of G rather than the distribution of $\hat{\theta}_j$'s or θ_j^*'s

EB Application: School Value-Added

- Next, illustrate/extend what we've learned in an application to estimating school value-added in Boston (Angrist, Hull, Pathak and Walters 2017)
- Suppose each student attends one of j schools, and let $Y_i(j)$ denote student i's potential academic achievement if s/he attends school j
- Simple additive model for potential outcomes:

$$Y_i(j) = \theta_j + a_i$$

- Here θ_j is the causal **value-added** of school j
- a_i represents unobserved student heterogeneity (family background, ability, etc.). Normalize $E[a_i] = 0$
- Constant effects model: $\theta_j - \theta_k$ is the effect of moving any student from school k to school j

Questions About Schools

- Several possible questions of interest in this setting
- Might be interested in the value-added of a particular school, e.g. θ_1
  - How good is my neighborhood school?
- Might be interested in features of the *distribution* of θ_j's across schools
  - How much does school quality vary?
- Might be interested in making a decision that depends on the θ_j's
  - Which school should my child attend? Which school(s) should be closed or expanded?
- EB methods are useful for answering each of these questions

VAM Regression

- Letting D_{ij} indicate attendance at j, observed outcome is:

$$Y_i = \sum_j \theta_j D_{ij} + a_i$$

- Project a_i on a vector of covariates X_i (e.g. demographics and lagged achievement):

$$Y_i = \sum_j \theta_j D_{ij} + X_i' \beta + \epsilon_i$$

- Here $E[X_i \epsilon_i] = 0$ by definition of β
- Suppose we have selection-on-observables: additive control for X_i captures all selection bias, so $E[D_{ij} \epsilon_i] = 0 \ \forall j$
- Then ordinary least squares (OLS) regression recovers the parameters of this value-added model (VAM)

Empirical Bayes for School Value-Added

- OLS VAM estimation yields estimates $\hat{\theta}_j$ and standard errors s_j (EB Step 1)
  - Assume $\hat{\theta}_j|\theta_j, s_j \sim N(\theta_j, s_j^2)$
  - Think of this as an asymptotic approximation with a growing number of students per school
- Second level of the hierarchy posits a mixing distribution for value-added:

$$\theta_j|s_j \sim N(\mu_\theta, \sigma_\theta^2)$$

- We can then apply the rest of the EB recipe
  - Step 2: Use $(\hat{\theta}_j, s_j)$ to estimate hyperparameters $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$
  - Step 3: Form linear shrinkage value-added estimates $\hat{\theta}_j^*$

VAM Standard Deviations for Boston Middle Schools (Sixth Grade Math)

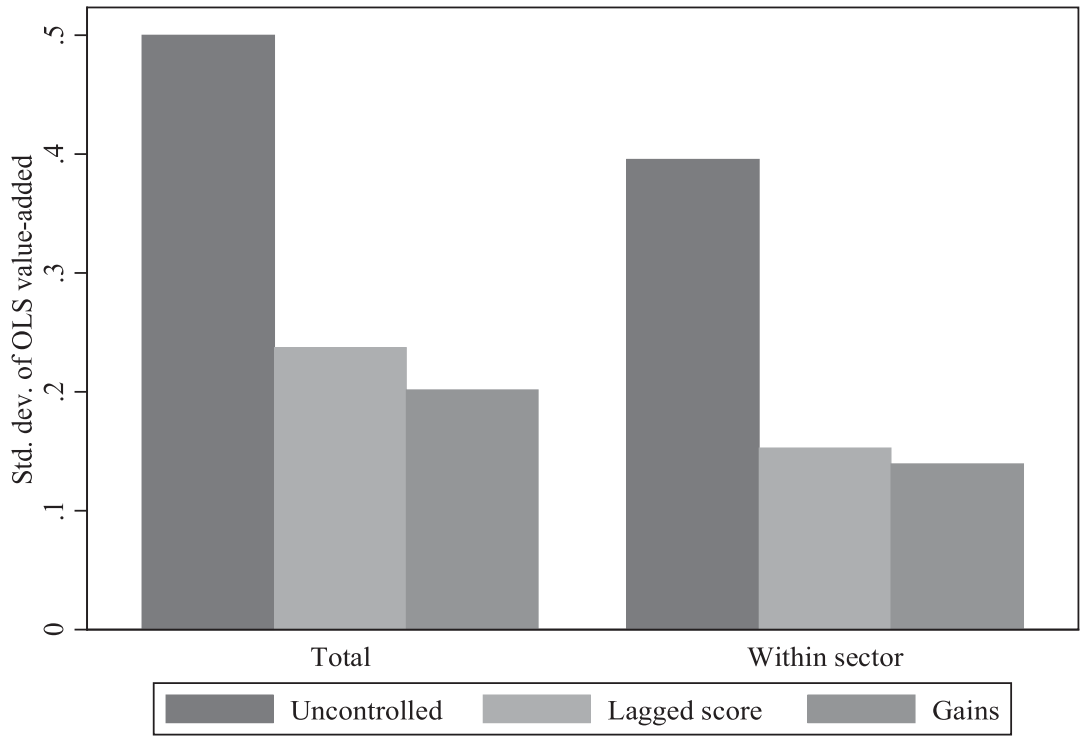
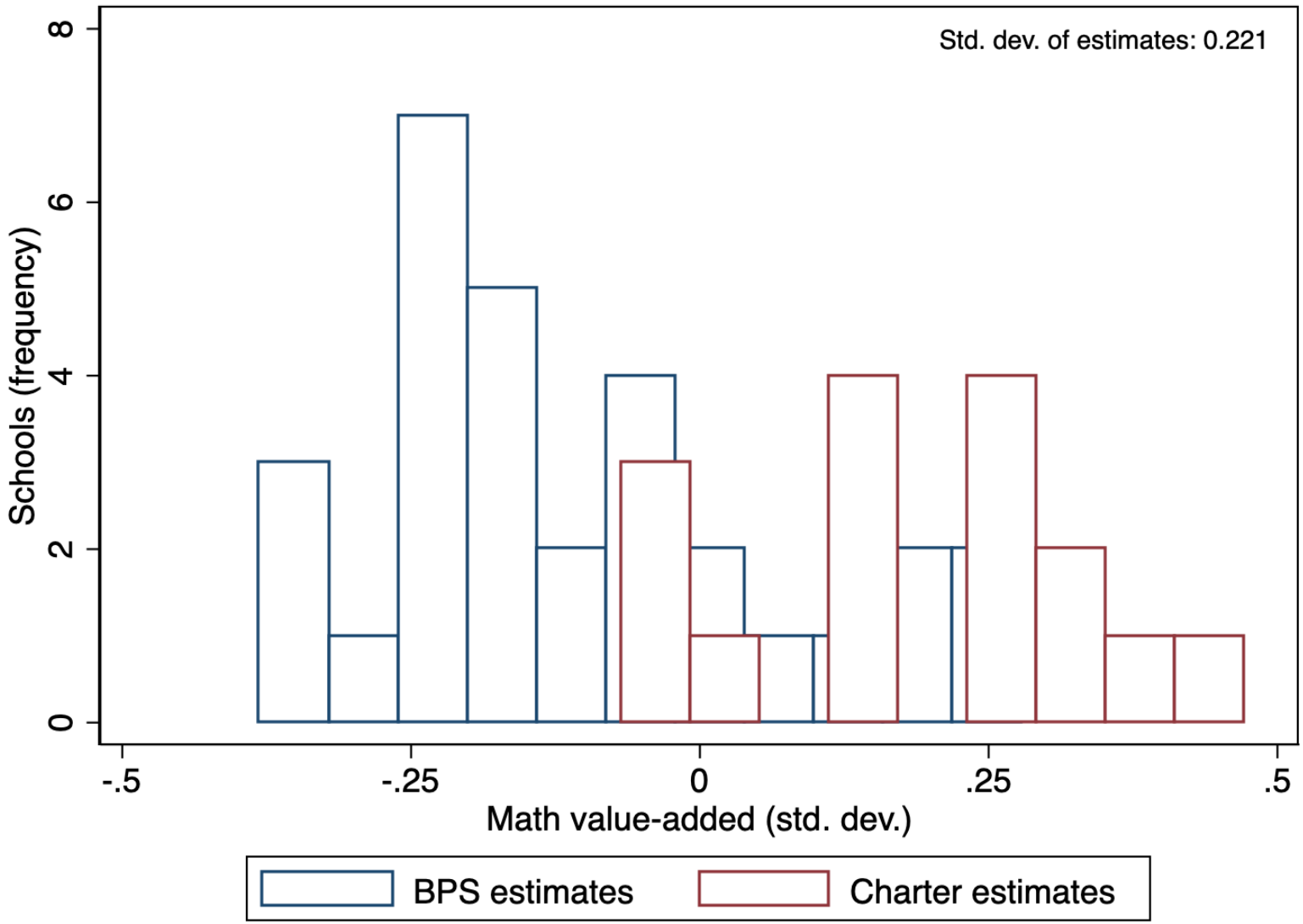


FIGURE I

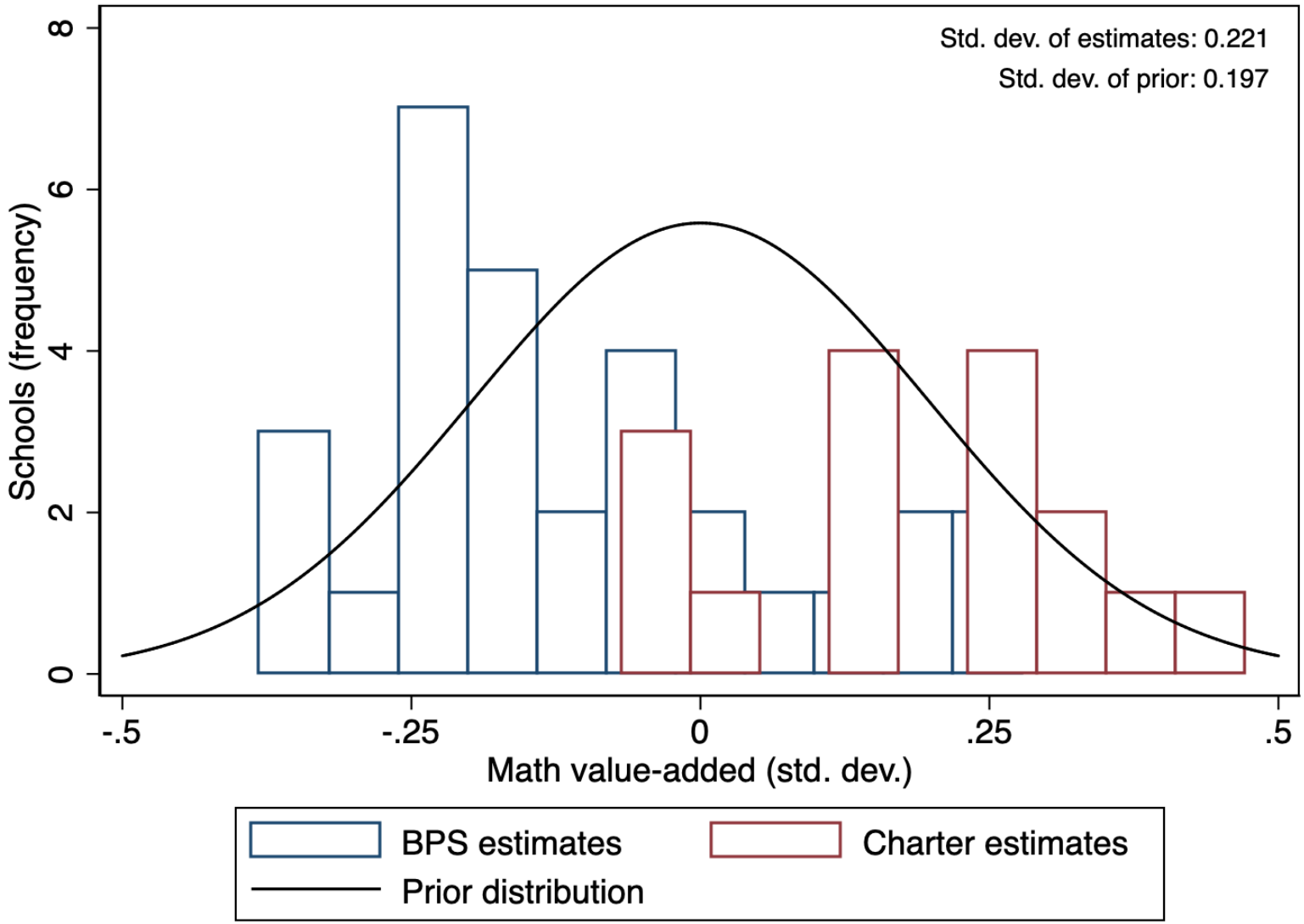
Standard Deviations of School Effects from OLS Value-Added Models

This figure compares standard deviations of school effects from alternative OLS value-added models. The notes to Table III describe the controls included in the lagged score and gains models; the uncontrolled model includes only year effects. The variance of OLS value-added is obtained by subtracting the average squared standard error from the sample variance of value-added estimates. Within-sector variances are obtained by first regressing value-added estimates on charter and pilot dummies, then subtracting the average squared standard error from the sample variance of residuals.

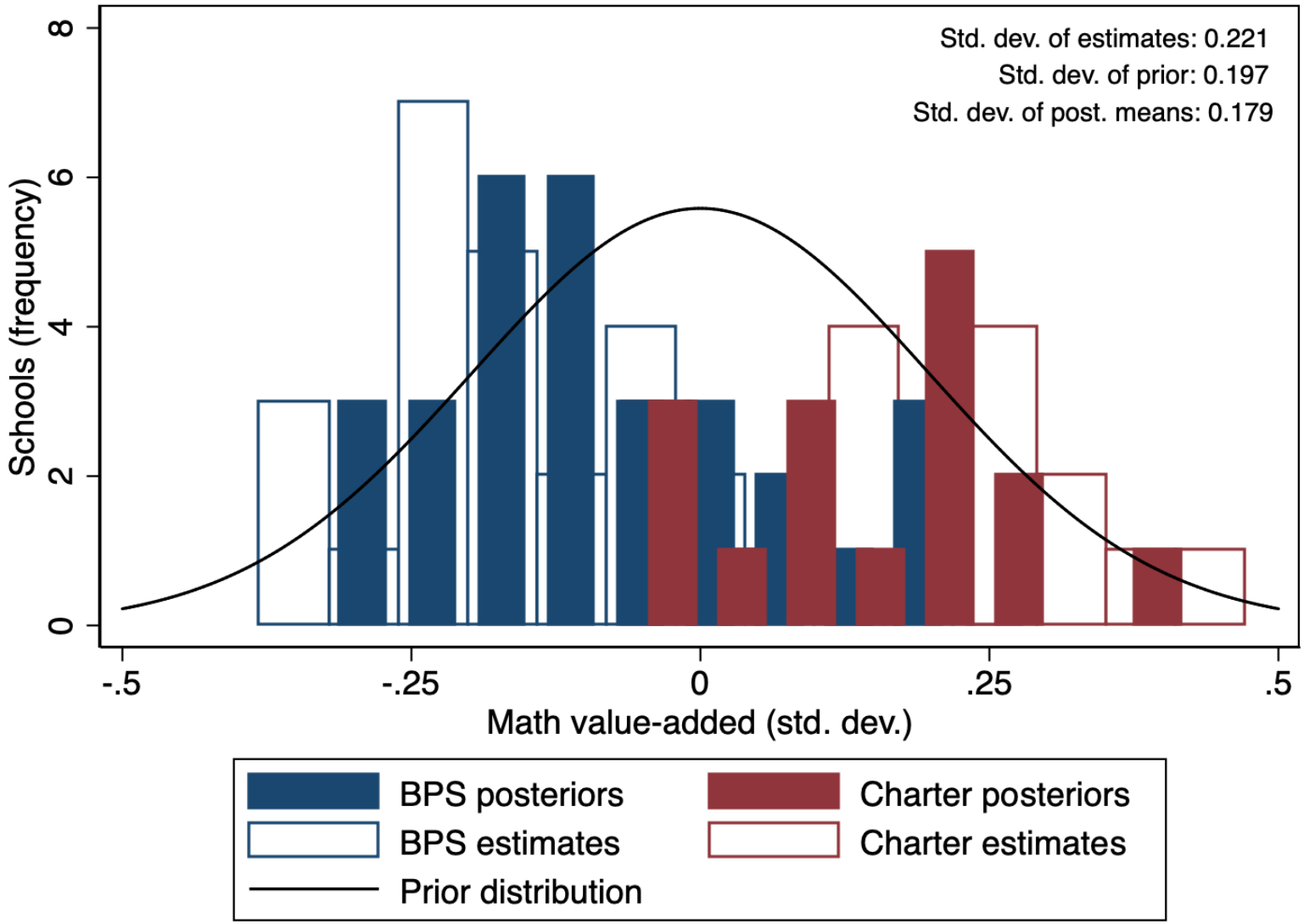
Histogram of Lagged Score VAM Estimates for Boston (Sixth Grade Math, 2014)



Prior Distribution Pooling Sectors



Posterior Means Pooling Sectors



Incorporating Covariates

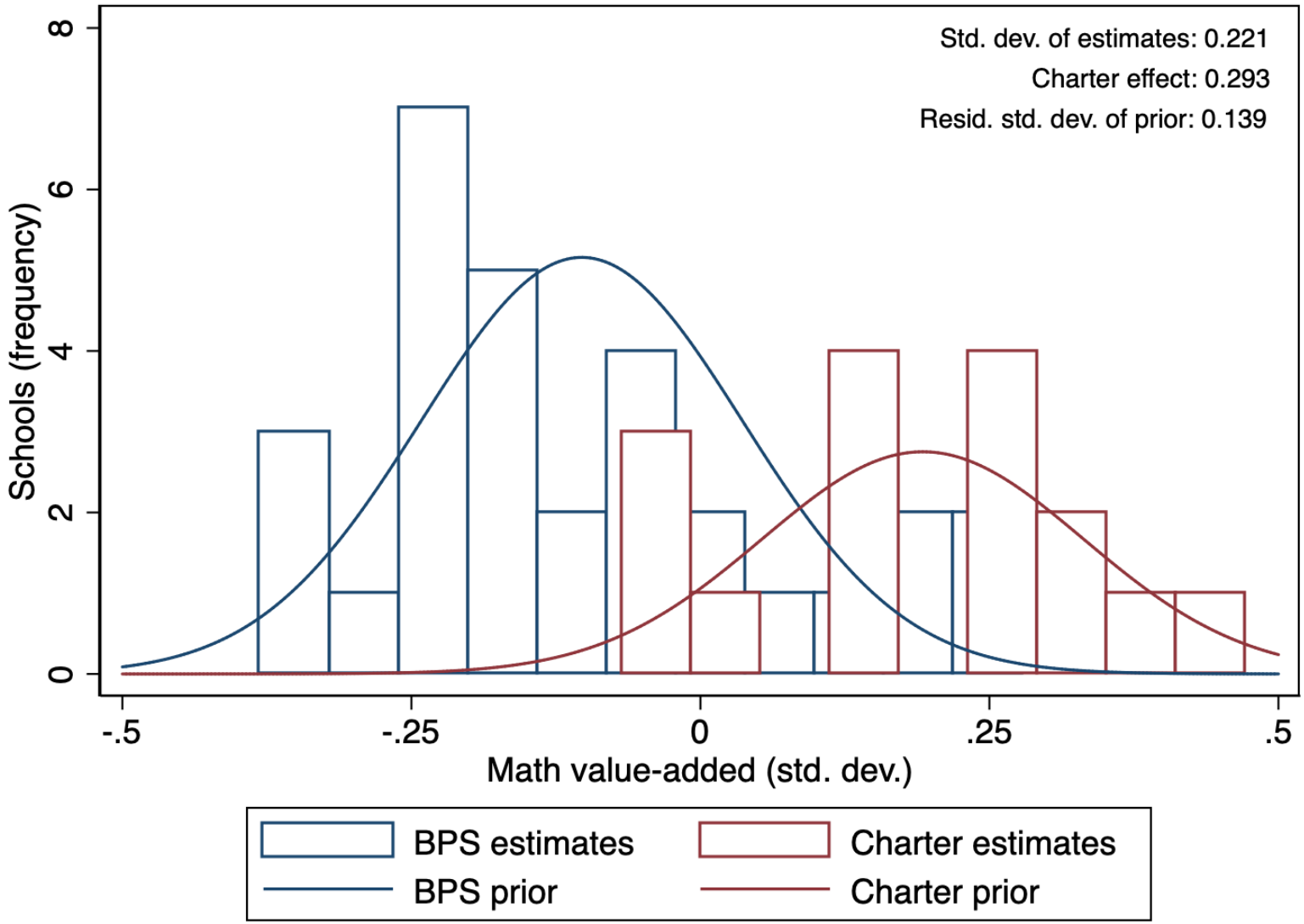
- It is often natural to build observed covariates into EB estimates
  - Learning from the experience of *which* others?
  - Mixed/multi-level models: allow θ_j's to depend on school-level characteristics (Raudenbush and Bryk, 1986)
- Model for G conditional on a vector of characteristics C_j, e.g. charter sector indicator:

$$\theta_j | s_j, C_j \sim N\left(C_j'\mu, \sigma_r^2\right)$$

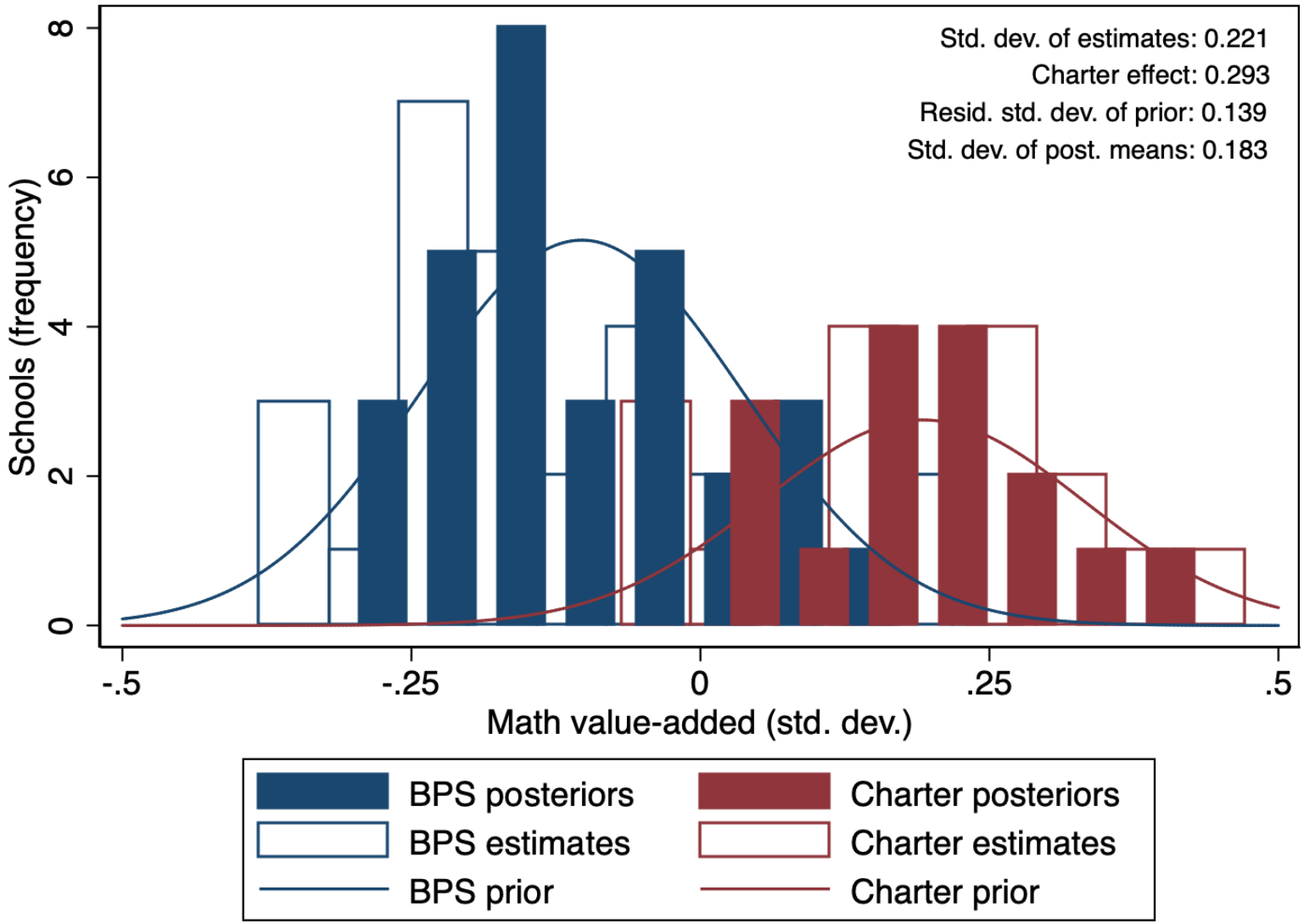
- Estimate μ from regression of $\hat{\theta}_j$ on C_j; deconvolve residuals $\hat{r}_j = \hat{\theta}_j - C_j'\hat{\mu}$ to estimate σ_r^2
- Resulting EB posterior shrinks $\hat{\theta}_j$ toward estimated linear index:

$$\hat{\theta}_j^* = \left(\frac{\hat{\sigma}_r^2}{\hat{\sigma}_r^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\hat{\sigma}_r^2 + s_j^2}\right)C_j'\hat{\mu}$$

Prior with Charter Sector Location Shift



Posteriors Shrinking Toward Sector Means



Combining Estimators

- EB framework extends naturally to cases where we have multiple estimates of the same parameter, some possibly biased
- Changing notation, let $\hat{\alpha}_j$ denote OLS estimate for school j, and suppose selection-on-observables fails, represented by bias parameter b_j:

$$\hat{\alpha}_j|\theta_j, b_j, s_{j\alpha} \sim N\left(\theta_j + b_j, s_{j\alpha}^2\right)$$

- Suppose we also have a noisy but (asymptotically) unbiased estimate $\hat{\theta}_j$, e.g. IV estimate from randomized lottery :

$$\hat{\theta}_j|\theta_j, b_j, s_{j\theta} \sim N(\theta_j, s_{j\theta}^2)$$

- Suppose a Hausman test rejects OLS = IV. Should we throw away OLS?

EB for Bias Correction

$$\hat{\alpha}_j|\theta_j, b_j, s_{j\alpha} \sim N\left(\theta_j + b_j, s_{j\alpha}^2\right),\ \hat{\theta}_j|\theta_j, b_j, s_{j\theta} \sim N(\theta_j, s_{j\theta}^2)$$

- We can use the ensemble $\{\hat{\alpha}_j, \hat{\theta}_j\}_{j=1}^J$ to estimate the joint distribution of truth and bias
- EB "hybrid" posterior $\hat{\theta}_j^* = E_{\hat{G}}[\theta_j|\hat{\theta}_j, \hat{\alpha}_j]$ trades off bias and variance to minimize MSE
- In special case with $s_{j,\alpha}^2 \approx 0$ we have

$$\hat{\theta}_j^* = \hat{\tau}_\theta \hat{\theta}_j + (1 - \hat{\tau}_\theta)\left(\hat{r}_\alpha(\hat{\alpha}_j - \hat{\mu}_b) + (1 - \hat{r}_\alpha)\hat{\mu}_\theta\right)$$

- Here $\tau_\theta = \frac{\sigma_\theta^2(1-R^2)}{\sigma_\theta^2(1-R^2)+s_{j\theta}^2}$, R^2 is R-squared from regression of θ_j on α_j, and r_α is slope coefficient from this regression (aka reliability of OLS)
- Angrist et al. (2017, forthcoming) generalize to underidentified case; see also Chetty and Hendren (2018)

MSE Improvements from Lottery-based Hybrid Estimates

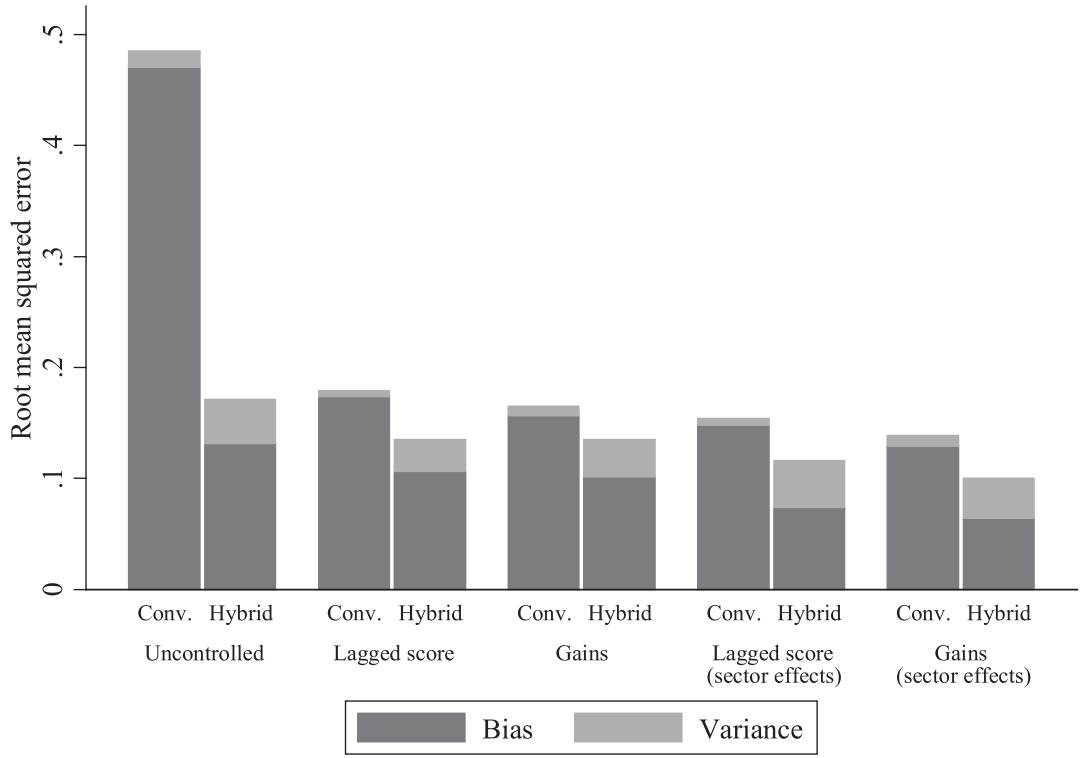


FIGURE VI

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.

Shrinkage and Regression

- Suppose we want to know how θ_j varies with a covariate C_j, e.g., charter status
- In other words, we'd like to know the regression of θ_j on C_j, given by $Cov(\theta_j, C_j)/Var(C_j)$
- Two options:
  1. Regress noisy $\hat{\theta}_j$ on C_j
  2. Regress shrunk $\hat{\theta}_j^*$ on C_j
- Which of these approaches recovers the regression of θ_j on C_j (if any)?

Shrinkage on the Left Causes Bias

- If noise $\hat{\theta}_j - \theta_j$ is independent of C_j, a regression of $\hat{\theta}_j$ on C_j gives the right answer:

$$\frac{Cov(\hat{\theta}_j, C_j)}{Var(C_j)} = \frac{Cov(\theta_j, C_j)}{Var(C_j)} + \frac{Cov(\hat{\theta}_j - \theta_j, C_j)}{Var(C_j)} = \frac{Cov(\theta_j, C_j)}{Var(C_j)}$$

- This is a consequence of the fact that classical measurement error on the left-hand side of a regression yields no bias

- In contrast, shrinkage on the left leads to bias. With a common shrinkage factor $\lambda = \sigma_\theta^2/(\sigma_\theta^2 + s^2)$, we have

$$\frac{Cov(\theta_j^*, C_j)}{Var(C_j)} = \frac{Cov(\lambda\hat{\theta}_j, C_j)}{Var(C_j)} = \lambda\frac{Cov(\hat{\theta}_j, C_j)}{Var(C_j)} = \lambda\frac{Cov(\theta_j, C_j)}{Var(C_j)}$$

Shrinkage and Regression

- What if we wanted to know how some variable W_j varies with θ_j?
- In other words, we'd like to know the regression of W_j on θ_j, given by $Cov(W_j, \theta_j)/Var(\theta_j)$
- Two options:
  1. Regress W_j on noisy $\hat{\theta}_j$
  2. Regress W_j on shrunk $\hat{\theta}_j^*$
- Which of these approaches recovers the regression of W_j on θ_j (if any)?

Shrinkage on the Right Fixes Bias

- When θ_j is on the right, using the unbiased estimate yields attenuation bias:

$$\frac{Cov(W_j, \hat{\theta}_j)}{Var(\hat{\theta}_j)} = \frac{Cov(W_j, \theta_j) + Cov(W_j, \hat{\theta}_j - \theta_j)}{Var(\theta_j) + Var(\hat{\theta}_j - \theta_j)} = \lambda \frac{Cov(W_j, \theta_j)}{Var(\theta_j)}$$

- In contrast, shrinkage of the regressor gives the right answer:

$$\frac{Cov(W_j, \lambda\hat{\theta}_j)}{Var(\lambda\hat{\theta}_j)} = \frac{\lambda Cov(W_j, \hat{\theta}_j)}{\lambda^2 Var(\hat{\theta}_j)} = \frac{1}{\lambda} \frac{Cov(W_j, \hat{\theta}_j)}{Var(\hat{\theta}_j)} = \frac{Cov(W_j, \theta_j)}{Var(\theta_j)}$$

- This is a version of errors-in-variables regression – use estimated signal-to-noise ratio to correct attenuation bias
- Rule of thumb: Put unbiased estimates on the left, and shrunk posteriors on the right

EB Decision Rules

- EB posterior means deliver estimates with low MSE
- We often have goals other than minimizing MSE
- Example: Suppose we want to close schools with value-added below a cutoff c (Gu and Koenker, 2023)
- Loss function for decision $\delta_j \in \{0, 1\}$:

$$\mathcal{L}(\theta_j, \delta_j) = \delta_j 1\{\theta_j > c\} + (1 - \delta_j) 1\{\theta_j \leq c\} \kappa$$

- Cost 1 of mistakenly closing high-performing school; cost κ of failing to close low-performing school

Minimizing Risk

- We do not observe the θ_j's, so we must choose a decision rule $\delta(\hat{\theta}_j, s_j)$ using noisy estimates and standard errors
- With J schools, the **risk** (expected loss) of decision rule δ is given by:

$$\mathcal{R}_\delta = E\left[\sum_j \mathcal{L}(\theta_j, \delta(\hat{\theta}_j, s_j))\right]$$

$$= \sum_j \int\int \mathcal{L}(\theta, \delta(\hat{\theta}, s_j)) \frac{1}{s_j} \phi\left(\frac{\hat{\theta} - \theta}{s_j}\right) d\hat{\theta} dG(\theta)$$

- Risk-minimizing decision rule:

$$\delta^* = \arg\min_{\delta \in \mathcal{D}} \sum_j \int\int \mathcal{L}(\theta, \delta(\hat{\theta}, s_j)) \frac{1}{s_j} \phi\left(\frac{\hat{\theta} - \theta}{s_j}\right) d\hat{\theta} dG(\theta)$$

Optimal Decision Rule

- With type I/type II loss function, solution is to select schools with sufficiently high posterior probability of value-added below c:

$$\delta^*(\hat{\theta}_j, s_j) = 1\left\{\text{Pr}_G\left[\theta_j < c|\hat{\theta}_j, s_j\right] \geq \frac{1}{1+\kappa}\right\}$$

- This means we should select based on posterior $(1/(1+\kappa))$ quantile rather than posterior mean
- In normal/normal model:

$$\delta^*(\hat{\theta}_j, s_j) = 1\left\{\left(\frac{\sigma^2_\theta}{\sigma^2_\theta+s^2_j}\right)\hat{\theta}_j + \left(\frac{s^2_j}{\sigma^2_\theta+s^2_j}\right)\mu_\theta + \sqrt{\frac{\sigma^2_\theta s^2_j}{\sigma^2_\theta+s^2_j}}\Phi^{-1}\left(\frac{1}{1+\kappa}\right) \leq c\right\}$$

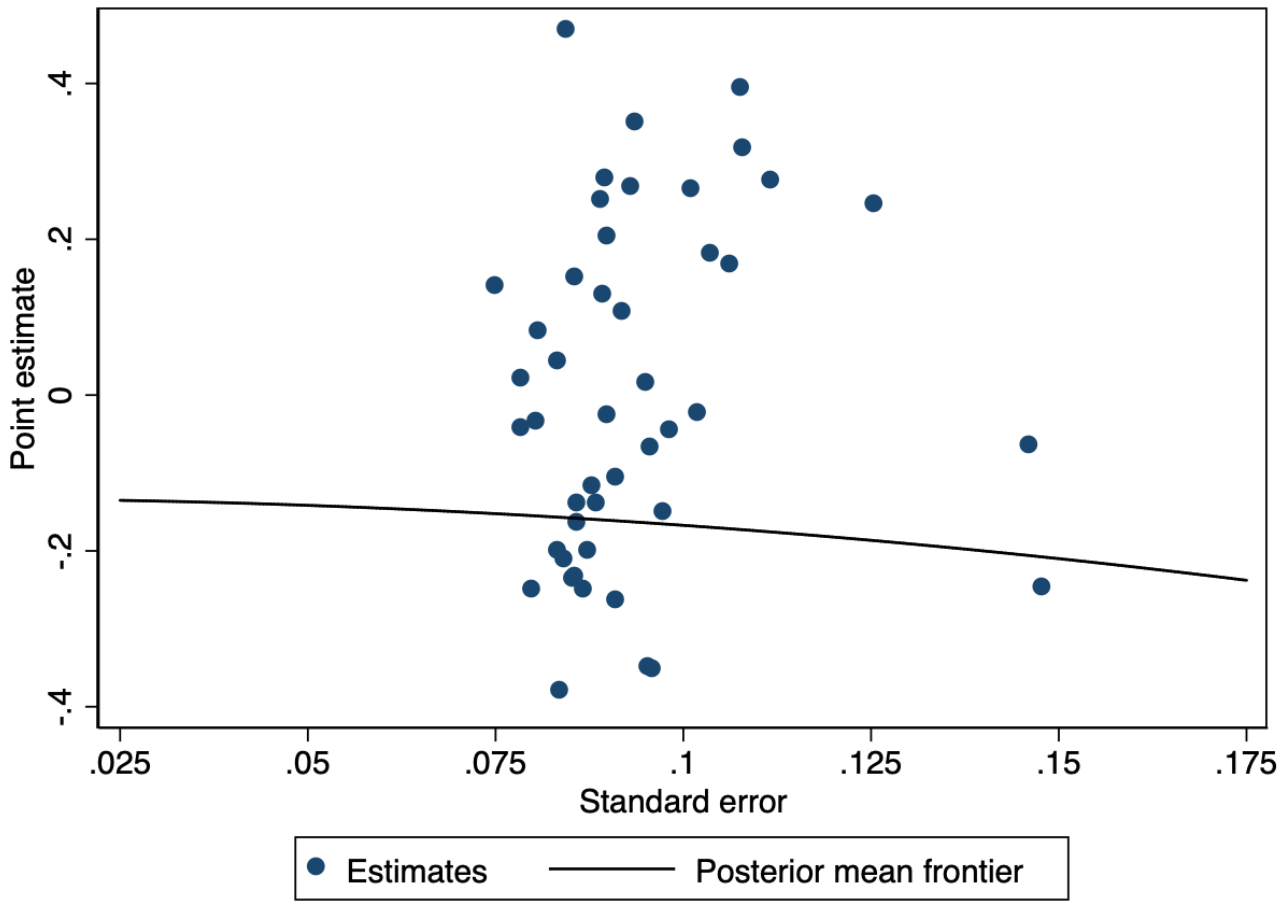
- EB decision rule $\hat{\delta}^*(\hat{\theta}_j, s_j)$ plugs in estimates $(\hat{\mu}_\theta, \hat{\sigma}_\theta)$

EB Decision Rule: Discussion

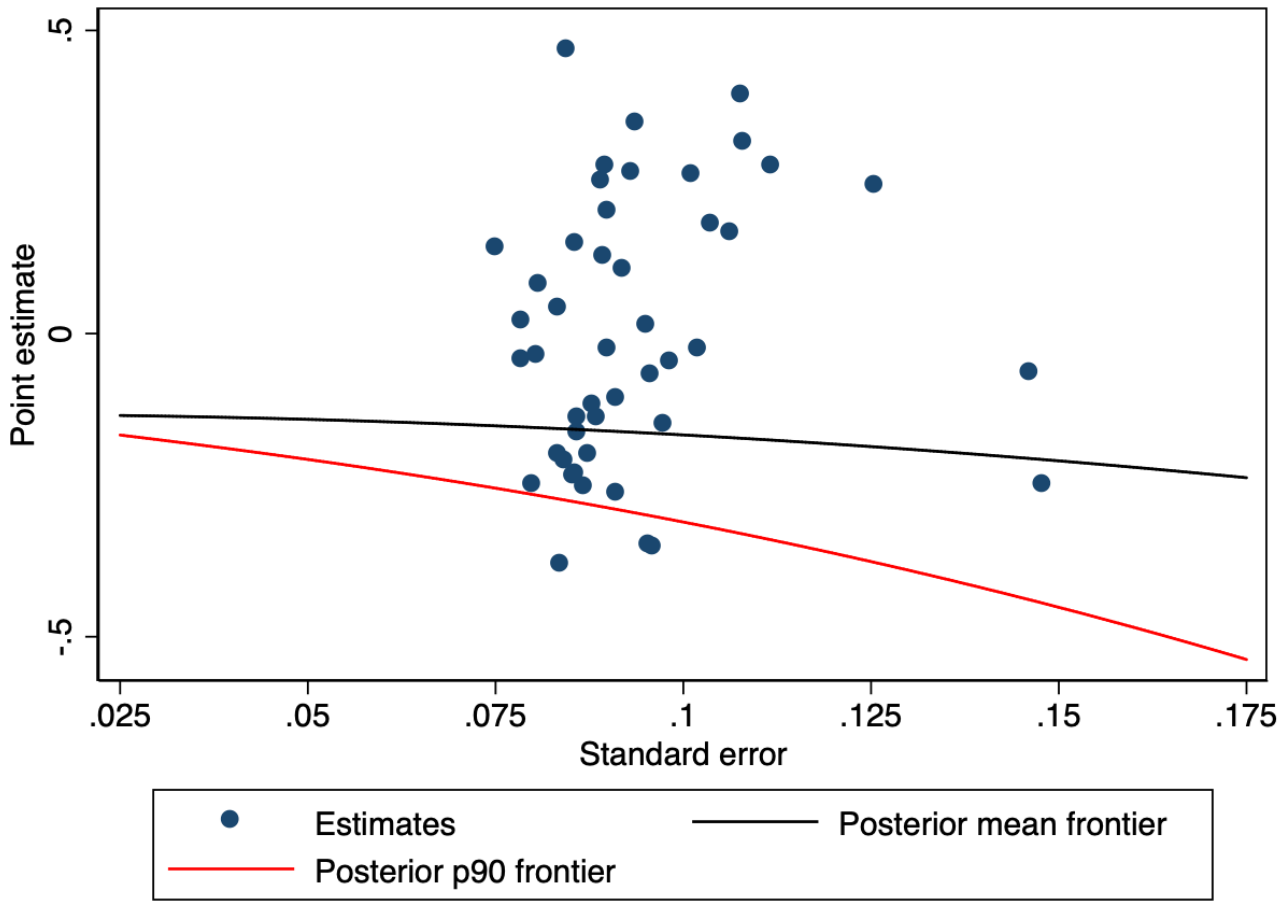
$$\hat{\delta}^*(\hat{\theta}_j, s_j) = 1\left\{\left(\frac{\hat{\sigma}^2_\theta}{\hat{\sigma}^2_\theta + s^2_j}\right)\hat{\theta}_j + \left(\frac{s^2_j}{\hat{\sigma}^2_\theta + s^2_j}\right)\hat{\mu}_\theta + \sqrt{\frac{\hat{\sigma}^2_\theta s^2_j}{\hat{\sigma}^2_\theta + s^2_j}}\Phi^{-1}\left(\frac{1}{1+\kappa}\right) \leq c\right\}$$

- Adjust posterior mean by $\sqrt{\sigma^2_\theta s^2_j/(\sigma^2_\theta + s^2_j)}\Phi^{-1}\left(1/(1+\kappa)\right)$, which may be positive or negative depending on costs of type I vs. II errors
  - Punish schools with larger standard errors if we care more about failing to close low performers
  - Reward schools with larger standard errors if we care more about mistakenly closing high performers
- Note that schools with the same posterior mean will be treated differently based on standard errors – raises potential horizontal equity issues
- General principle: different objectives call for using different features of posterior for decision-making
  - It's important to state your loss function!

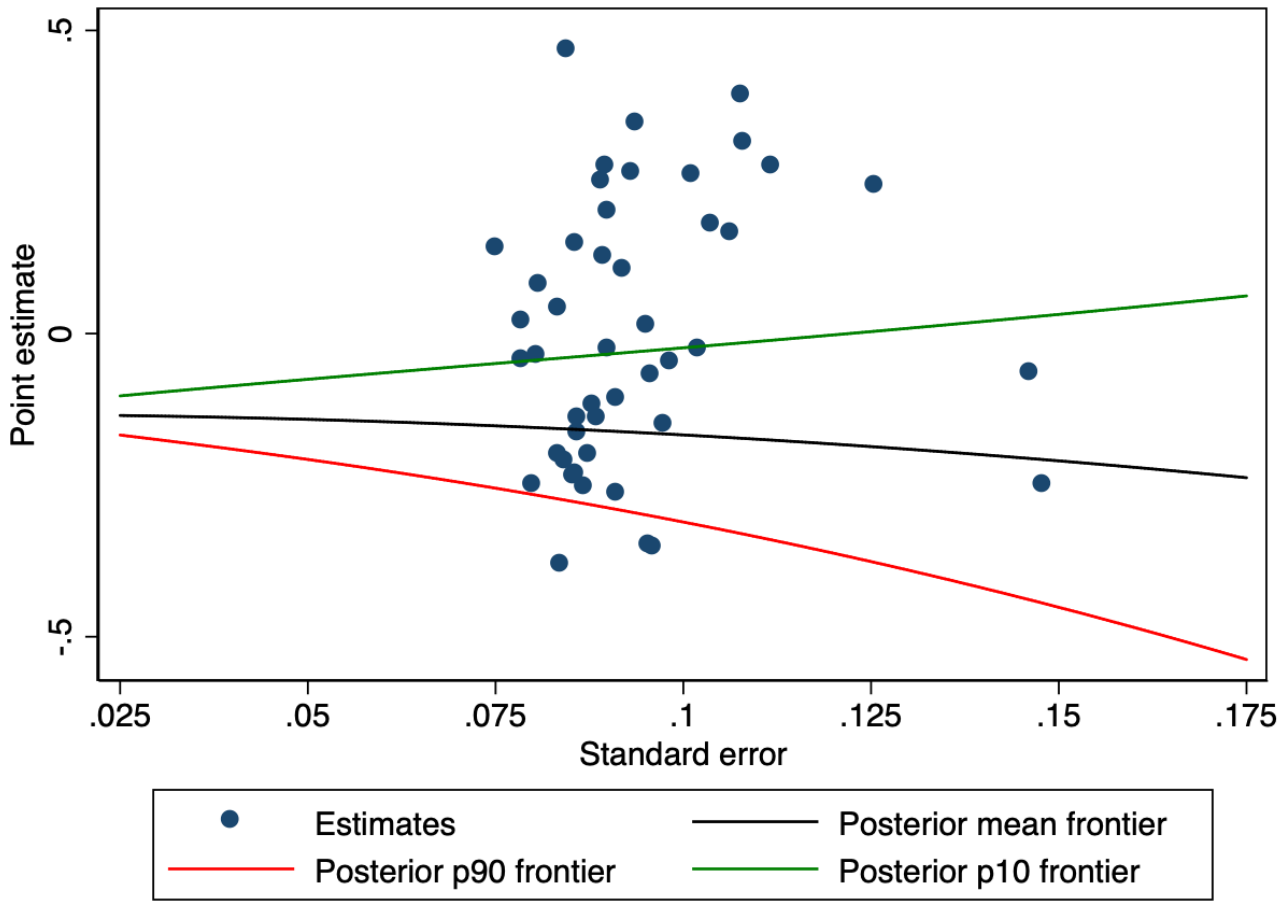
Posterior Mean Decision Frontier: $c = \mu_\theta + \sigma_\theta \Phi^{-1}(0.25)$



Posterior 90th Percentile Rule Rewards Large s_j



Posterior 10th Percentile Rule Penalizes Large s_j



Individualized Treatment Effect Predictions

- Another increasingly common use case: individualized treatment effect predictions or treatment recommendations
- Consider a randomized controlled trial of a binary treatment $T_i \in \{0, 1\}$
- Potential outcomes $Y_i(1)$ and $Y_i(0)$ describe i's outcomes in the treated and untreated states
- Suppose each individual also belongs to one of G subgroups, indicated by $G_i \in \{1, ..., G\}$
- Randomization $\implies$ treatment is independent of potential outcomes within each group: $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i | G_i$
- Treatment/control comparisons in each group therefore identify conditional average treatment effects (CATEs) for each group:

$$E[Y_i | T_i = 1, G_i = g] - E[Y_i | T_i = 0, G_i = g] = E[Y_i(1) - Y_i(0) | G_i = g] \equiv \Delta_g$$

Individualized Treatment Effect Predictions

- Treatment/control contrasts yield estimates $\hat{\Delta}_g$ and standard errors s_g
- If asked to provide a treatment effect prediction for each group, one option is to report $\hat{\Delta}_g$. Can we do better?
- Linear shrinkage estimate of CATE for group g:

$$\hat{\Delta}_g^* = \left(\frac{\hat{\sigma}_\Delta^2}{\hat{\sigma}_\Delta^2 + s_g^2} \right) \hat{\Delta}_g + \left(\frac{s_g^2}{\hat{\sigma}_\Delta^2 + s_g^2} \right) \hat{\mu}_\Delta$$

- Shrinkage reduces MSE by "borrowing strength" from overall average treatment effect estimate $\hat{\mu}_\Delta$, avoiding excess weight on small/noisy subgroups
- If we want to recommend the treatment with higher mean potential outcome for each group, may want to base recommendations on posterior probability of positive CATE
  - With normal prior G, $Pr_{\hat{G}}[\Delta_g > 0 | \hat{\Delta}_g, s_g] = \Phi \left(\frac{\hat{\Delta}_g^*}{[\hat{\sigma}_\Delta^2 s_g^2 / (\hat{\sigma}_\Delta^2 + s_g^2)]} \right)$

EB and Machine Learning

- EB methods are closely related to **machine learning** (ML) approaches
- ML refers to a suite of tools for model selection/penalization in settings with many predictors
  - LASSO, ridge regression, random forests, neural nets, transformers, etc.
- Goal of ML is to reduce overfitting in finite samples, thereby obtaining a better estimate of some conditional distribution function
  - For example, estimate $E[Y_i|X_i]$ when number of regressors in X_i is bigger than sample size
- Sounds a lot like using EB shrinkage to reduce variance and improve MSE....

EB and Machine Learning

- To draw connections between EB and ML, return to parametric normal/normal model with N observations per group:

$$Y_{ij} = \theta_j + \varepsilon_{ij}$$

$$\varepsilon_{ij}|\theta_j \sim N(0, \sigma_\epsilon^2)$$

$$\theta_j \sim N(0, \sigma_\theta^2)$$

- Unbiased estimator $\hat{\theta}_j = \frac{1}{N}\sum_i Y_{ij}$ with sampling variance $s_j^2 = \sigma_\epsilon^2/N$
- Posterior distribution for θ_j is $N(\theta_j^*, V^*)$ with

$$\theta_j^* = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2/N}\right)\hat{\theta}_j, \;\; V^* = \frac{\sigma_\epsilon^2\sigma_\theta^2}{N\sigma_\theta^2 + \sigma_\epsilon^2}$$

EB and Machine Learning

- Posterior density for θ_j:

$$f(\theta_j|Y_{1j},....,Y_{Nj}) = \frac{\left[\prod_{i=1}^{N} \frac{1}{\sigma_\epsilon}\phi\left(\frac{Y_{ij}-\theta_j}{\sigma_\epsilon}\right)\right]\frac{1}{\sigma_\theta}\phi\left(\frac{\theta_j}{\sigma_\theta}\right)}{\int_{-\infty}^{\infty}\left[\prod_{i=1}^{N} \frac{1}{\sigma_\epsilon}\phi\left(\frac{Y_{ij}-\theta}{\sigma_\epsilon}\right)\right]\frac{1}{\sigma_\theta}\phi\left(\frac{\theta}{\sigma_\theta}\right)d\theta}$$

- Posterior distribution is normal $\implies$ posterior mean and mode coincide
- This implies posterior means maximize log posterior density:

$$(\theta_1^*,...,\theta_J^*) = \arg\max_{(\theta_1,...,\theta_J)} \sum_j \log f(\theta_j|Y_{1j}....Y_{Nj})$$

$$= \arg\max_{(\theta_1,...,\theta_J)} \sum_{j=1}^{J}\sum_{i=1}^{N} \log\phi\left(\frac{Y_{ij}-\theta_j}{\sigma_\epsilon}\right) + \sum_{j=1}^{J}\log\phi\left(\frac{\theta_j}{\sigma_\theta}\right) + cons$$

- Posterior mode is also known as a **maximum a posteriori** (MAP) estimate

EB and Machine Learning

- Plugging in normal density yields

$$(\theta_1^*, ..., \theta_J^*) = \arg \max_{(\theta_1,...,\theta_J)} -\sum_{j=1}^{J}\sum_{i=1}^{N} \frac{(Y_{ij} - \theta_j)^2}{2\sigma_\epsilon^2} - \sum_{j=1}^{J} \frac{\theta_j^2}{2\sigma_\theta^2}$$

$$= \arg \min_{(\theta_1,...,\theta_J)} \sum_{j=1}^{J}\sum_{i=1}^{N}(Y_{ij} - \theta_j)^2 + \frac{\sigma_\epsilon^2}{\sigma_\theta^2}\sum_{j=1}^{J}\theta_j^2$$

$$= \arg \min_{(\theta_1,...,\theta_J)} \sum_{j=1}^{J}\sum_{i=1}^{N}(Y_{ij} - \theta_j)^2 + \lambda p(\theta_1, ..., \theta_J)$$

- This is regularized least squares with an L2 (quadratic) penalty $p(\cdot)$, also known as **ridge regression**
- Empirical Bayes $\implies$ use the data to choose tuning parameter λ in penalty function (i.e. estimate $\sigma_\epsilon^2/\sigma_\theta^2$)

EB and Machine Learning

- ML penalization/regularization procedures often have an EB interpretation
  - Ridge regression estimates (L2 penalization) can be interpreted as posterior means from a model with normal priors
  - LASSO estimates (L1 penalization) can be interpreted as MAP estimates from a model with double exponential (Laplace) priors
- When doing model selection or penalization via ML, useful to think about implicit prior distribution and connection to loss function
- See Abadie and Kasy (2019) for analysis of the relative performance of common regularization approaches under various G's

Empirical Bayes vs. Full Bayes

- A fully Bayesian analysis would add a third level to the hierarchy: a **hyperprior** over the mixing distribution, with parameters chosen by the researcher rather than estimated
- For example, we might have a normal/normal model with a normal-inverse Gamma hyperprior:

$$\hat{\theta}_j|\theta_j, s_j \sim N(\theta_j, s_j^2)$$

$$\theta_j|s_j \sim N(\mu_\theta, \sigma_\theta^2)$$

$$(\mu_\theta, \sigma_\theta^2)|\mathbf{s} \sim N - \Gamma^{-1}(m, \lambda, \alpha, \beta)$$

- We would then choose values for $(m, \lambda, \alpha, \beta)$ and compute posterior distributions for μ_θ, σ_θ, and each θ_j
- Why should we opt for empirical Bayes rather than fully Bayesian methods?

EB vs. Full B

- Pros of full Bayes:
  - Internally consistent and coherent approach to updating beliefs about parameters (if you're a Bayesian). But if you're a frequentist....
  - EB posteriors do not account for estimation error in hyperparameters, so can overstate precision (though we can adjust for this)
  - In some cases hyperparameters are difficult to estimate, and smoothing via a hyperprior can help
- Cons of full Bayes:
  - Fully Bayesian estimation often requires simulation methods (Markov Chain Monte Carlo, MCMC), which are harder to implement and less transparent
  - Where do the parameters of the hyperprior come from? To the extent that these affect the estimates, why should we believe the results?
- EB estimates have desirable frequentist properties and are easier to understand – arguably less "harmful"
- If hyperparameters are estimated precisely, there won't be much difference

Robust/Non-parametric Empirical Bayes

- Our discussion so far has focused on parametric models with normality/independence assumptions
  - Though we've seen EB procedures can be robust to violations of these assumptions (James/Stein; linear shrinkage as best linear approximation)
- Next, consider robust/non-parametric EB approaches that relax these assumptions
  - Generalized variance estimation
  - Precision-dependence
  - Non-parametric deconvolution
  - Linear vs. non-parametric shrinkage
- Main example: study of employer-level labor market discrimination by Kline, Rose and Walters (2022)

Generalized Variance Estimation

- Suppose we are interested in a vector of J parameters $\Theta = (\theta_1, ..., \theta_J)'$
- We have estimates $\hat{\theta}_j$ collected in vector $\hat{\Theta} = (\hat{\theta}_1, ..., \hat{\theta}_J)'$
- Assume $\hat{\Theta}$ is an unbiased estimate of Θ: $E[\hat{\Theta}|\Theta] = \Theta$
- Matrix V describes the variance of noise in $\hat{\Theta}$:

$$V = E\left[(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)'|\Theta\right]$$

- With independence across units, V has s_j^2's on the diagonal and zeros elsewhere. In other cases, off-diagonal elements may not be zero

Correlation Across Units

- When should we worry about correlation in noise across units?
- Suppose we fit OLS VAMs for two outcomes, Y_i^1 and Y_i^2 (e.g. test scores and social skills):

$$Y_i^k = \sum_j \theta_j^k D_{ij} + X_i' \beta^k + \epsilon_i^k, \ k \in \{1, 2\}$$

- Here the full vector of estimates is $\hat{\Theta} = (\hat{\theta}_1^1, ..., \hat{\theta}_J^1, \hat{\theta}_1^2, ..., \hat{\theta}_J^2)'$
- $\hat{\theta}_j^1$ and $\hat{\theta}_j^2$ are estimated from the same data so likely highly correlated – V is not diagonal
- We will need to account for this correlation if we want to study relationships between parameters across equations (e.g. covariance between θ_j^1 and θ_j^2)

Quadratic Forms

- Suppose we are interested in a quadratic form involving Θ:

$$\delta = \Theta' A \Theta$$

- A is some known $J \times J$ matrix
- For example, we might have $A = J^{-1}\left(I_J - J^{-1}\iota_J\iota_J'\right)$, with I_J the $J \times J$ identity and ι_J a $J \times 1$ vector of 1's
- With this choice of A, δ is the variance of θ_j's across units:

$$\delta = \frac{1}{J}\sum_{j=1}^{J}(\theta_j - \bar{\theta})^2, \ \bar{\theta} = \frac{1}{J}\sum_{j=1}^{J}\theta_j$$

- In the two-equation example, we can obtain $Cov(\theta_j^1, \theta_j^2)$ by choosing $A = \begin{bmatrix} 0 & J^{-1}\left(I_J - J^{-1}\iota_J\iota_J'\right) \\ 0 & 0 \end{bmatrix}$

Plug-in Estimator

- Plug-in estimator of quadratic form δ:

$$\hat{\delta} = \hat{\Theta}' A \hat{\Theta}$$

- This estimator is biased: $E[\hat{\delta}|\Theta] \neq \delta$
- Intuitively, $\hat{\theta}_j^2$ is an upward-biased estimate of θ_j^2 due to noise
  - Even if all θ_j's were equal, we'd get some variation in the $\hat{\theta}_j$'s by chance
  - Plug-in estimator does not account for the contribution of sampling error
  - Generalization of the idea that variance of $\hat{\theta}_j$'s is too big relative to true θ_j's

Bias of Plug-in Estimator

- Formally, expectation of the plug-in estimator is

$$E[\hat{\delta}|\Theta] = E[\hat{\Theta}'A\hat{\Theta}|\Theta]$$

$$= E\left[\left(\Theta + (\hat{\Theta} - \Theta)\right)' A \left(\Theta + (\hat{\Theta} - \Theta)\right)|\Theta\right]$$

$$= \Theta'A\Theta + \Theta'AE\left[\hat{\Theta} - \Theta|\Theta\right] + E\left[\hat{\Theta} - \Theta|\Theta\right]' A\Theta + E\left[(\hat{\Theta} - \Theta)'A(\hat{\Theta} - \Theta)|\Theta\right]$$

$$= \delta + E\left[tr\left((\hat{\Theta} - \Theta)'A(\hat{\Theta} - \Theta)\right)|\Theta\right]$$

$$= \delta + tr\left(AE\left[(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)'|\Theta\right]\right)$$

$$= \delta + tr(AV).$$

- Bias of plug-in estimator is therefore $E[\hat{\delta} - \delta|\Theta] = tr\,(AV)$

Bias-Corrected Variance Estimation

- A bias-corrected estimator of δ using a variance estimate $\hat{V}$ is given by

$$\hat{\delta}_{BC} = \hat{\delta} - tr(A\hat{V}) = \hat{\Theta}' A \hat{\Theta} - tr(A\hat{V})$$

- Subtract off expected contribution of noise from naive plug-in estimator
  - Analogous to subtracting off average squared SE from sample variance of estimates in earlier example
- If $\hat{V}$ is an unbiased estimate of the variance V so that $E[\hat{V}] = V$, then $\hat{\delta}_{BC}$ is unbiased for δ
- Kline, Saggio and Sølvsten (KSS, 2020) propose leave-out approach to obtain finite-sample unbiased estimate of V

Precision-Weighting

- So far we have assumed that effect sizes θ_j are independent of sampling variance s_j^2 across units
- If θ_j and s_j^2 are independent, we may want to weight estimates of mean and/or variance of G to account for differences in noise:

$$\hat{\mu}_\theta = \sum_{j=1}^{J} w_j \hat{\theta}_j, \ \sum_{j=1}^{J} w_j = 1$$

- Precision-weighting uses weight proportional to $1/s_j^2$
- Optimal (minimum-variance) weight is proportional to $1/(\sigma_\theta^2 + s_j^2)$
  - Can form estimates of these weights based on first-step unweighted $\hat{\sigma}_\theta^2$ (analogous to FGLS)
  - Maximum likelihood jointly estimates $(\mu_\theta, \sigma_\theta^2)$ and optimal weights in one step (analogous to CUE)
- But precision-weighting changes the estimand if θ_j is correlated with s_j, which some find unappealing
  - An instance of general debate about weighting in econometrics

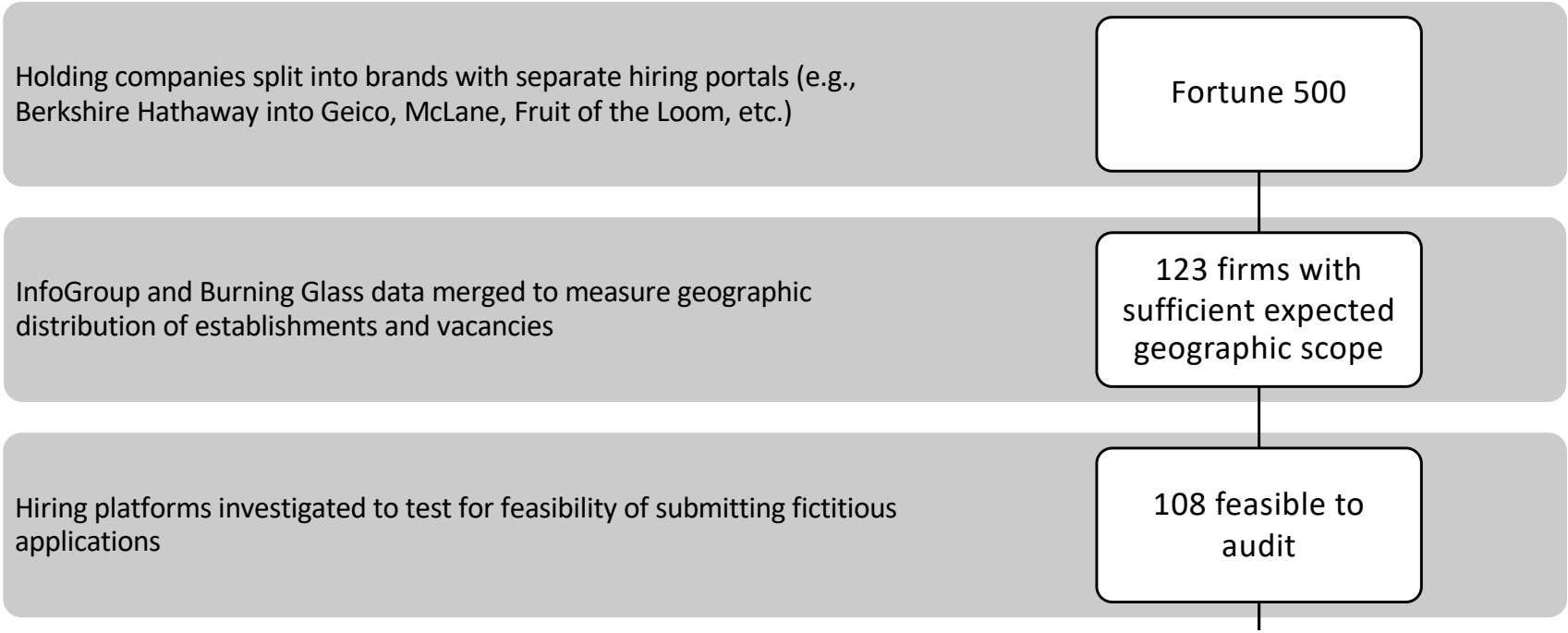
Interpreting Precision-Dependence

- Why would effect sizes θ_j be related to sampling variances s_j^2?
- Intuition: Recall that for sample mean, $s_j^2 = \sigma_j^2 / N_j$
  - Units with larger sample sizes N_j may have bigger/smaller effects
  - Units with more within-group variance σ_j^2 may have bigger/smaller effect
- In some applications, precision-dependence can be economically interesting
  - For example, teacher value-added literature finds that value-added increases with experience
  - This suggests that more experienced teachers will have higher θ_j and more data available to estimate it (lower s_j^2)
- Later we will return to approaches to dealing with precision-dependence

Application: Employer-level Labor Market Discrimination

- Example of robust/non-parametric EB techniques comes from a study of employment discrimination by Kline, Rose, and Walters (2022)
- Massive resume correspondence study sending applications to multiple establishments at large employers
  - 108 Fortune 500 firms
  - Up to 125 jobs per firm, each in a different county
  - 8 applications per job (stratified 4 Black/4 white)
- Following Bertrand and Mullainathan (2004), manipulate employer perceptions of race and sex using distinctive names

Sampling frame (I/II)

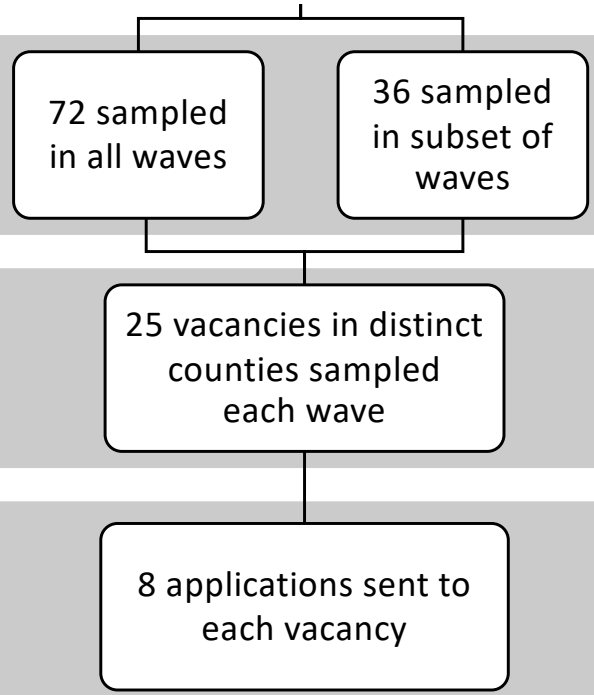


Sampling frame (II/II)

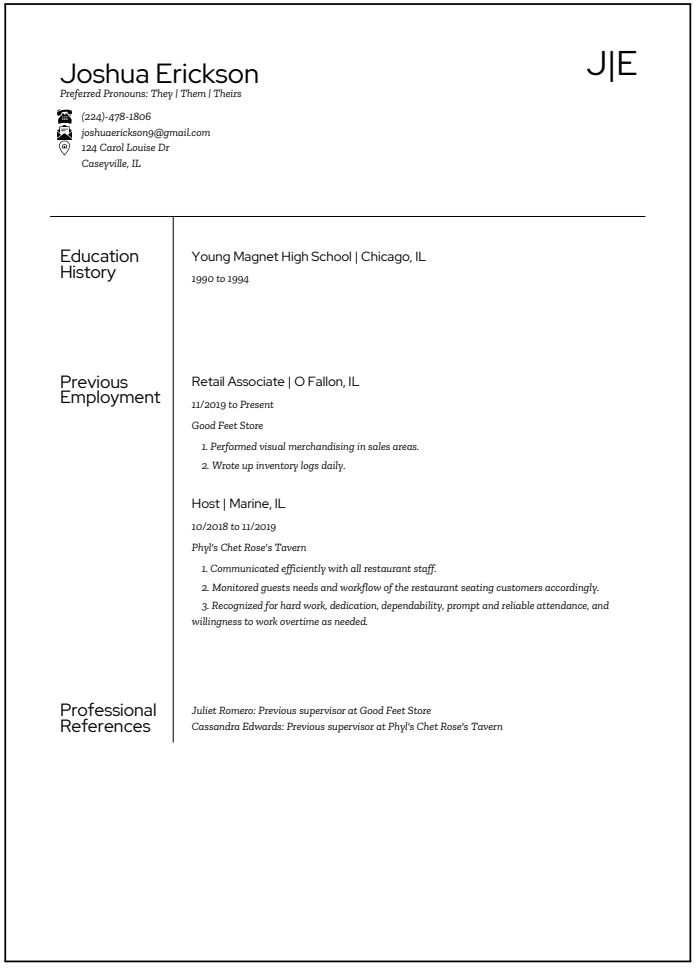
4 not sampled in wave 1 due to COVID interruption; 9 firms dropped before completion due to technological constraints; 19 added in wave 2 or later; 4 posted insufficient jobs to sample in all waves

Job sampled from universe of entry-level vacancies posted on each firm's hiring portal; most recently posted job prioritized

One pair of applications (1 black and 1 white name) sent every 1-2 days; gender (50% male), age (uniform age 20-60), gender identity (5% gender-neutral, 5% same-gender pronouns), and sexual orientation (10% LGBTQ student club, 10% other club) unconditionally randomly assigned



Example resumes



Joshua Erickson

Preferred Pronouns: They | Them | Theirs

(224)-478-1806
joshuaerickson9@gmail.com
124 Carol Louise Dr
Caseyville, IL

J|E

Education History

Young Magnet High School | Chicago, IL

1990 to 1994

Previous Employment

Retail Associate | O Fallon, IL

11/2019 to Present

Good Feet Store

1. *Performed visual merchandising in sales areas.*
2. *Wrote up inventory logs daily.*

Host | Marine, IL

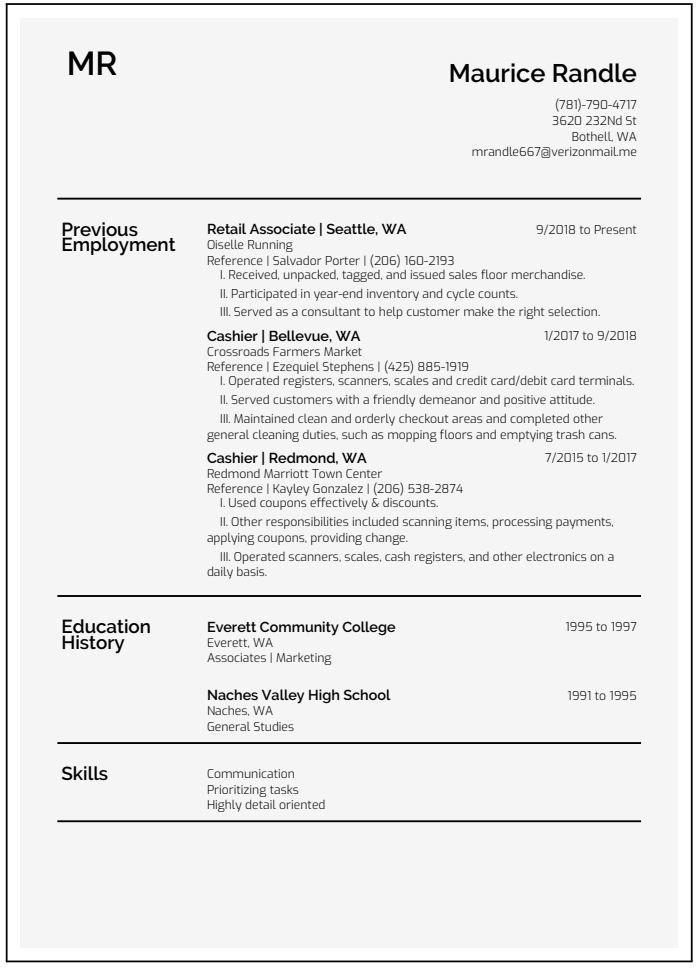
10/2018 to 11/2019

Phyl's Chet Rose's Tavern

1. *Communicated efficiently with all restaurant staff.*
2. *Monitored guests needs and workflow of the restaurant seating customers accordingly.*
3. *Recognized for hard work, dedication, dependability, prompt and reliable attendance, and willingness to work overtime as needed.*

Professional References

Juliet Romero: Previous supervisor at Good Feet Store
Cassandra Edwards: Previous supervisor at Phyl's Chet Rose's Tavern



MR

Maurice Randle

(781)-790-4717
3620 232Nd St
Bothell, WA
mrandle667@verizonmail.me

Previous Employment

Retail Associate | Seattle, WA 9/2018 to Present
Oiselle Running
Reference | Salvador Porter | (206) 160-2193

I. Received, unpacked, tagged, and issued sales floor merchandise.

II. Participated in year-end inventory and cycle counts.

III. Served as a consultant to help customer make the right selection.

Cashier | Bellevue, WA 1/2017 to 9/2018
Crossroads Farmers Market
Reference | Ezequiel Stephens | (425) 885-1919

I. Operated registers, scanners, scales and credit card/debit card terminals.

II. Served customers with a friendly demeanor and positive attitude.

III. Maintained clean and orderly checkout areas and completed other general cleaning duties, such as mopping floors and emptying trash cans.

Cashier | Redmond, WA 7/2015 to 1/2017
Redmond Marriott Town Center
Reference | Kayley Gonzalez | (206) 538-2874

I. Used coupons effectively & discounts.

II. Other responsibilities included scanning items, processing payments, applying coupons, providing change.

III. Operated scanners, scales, cash registers, and other electronics on a daily basis.

Education History

Everett Community College 1995 to 1997
Everett, WA
Associates | Marketing

Naches Valley High School 1991 to 1995
Naches, WA
General Studies

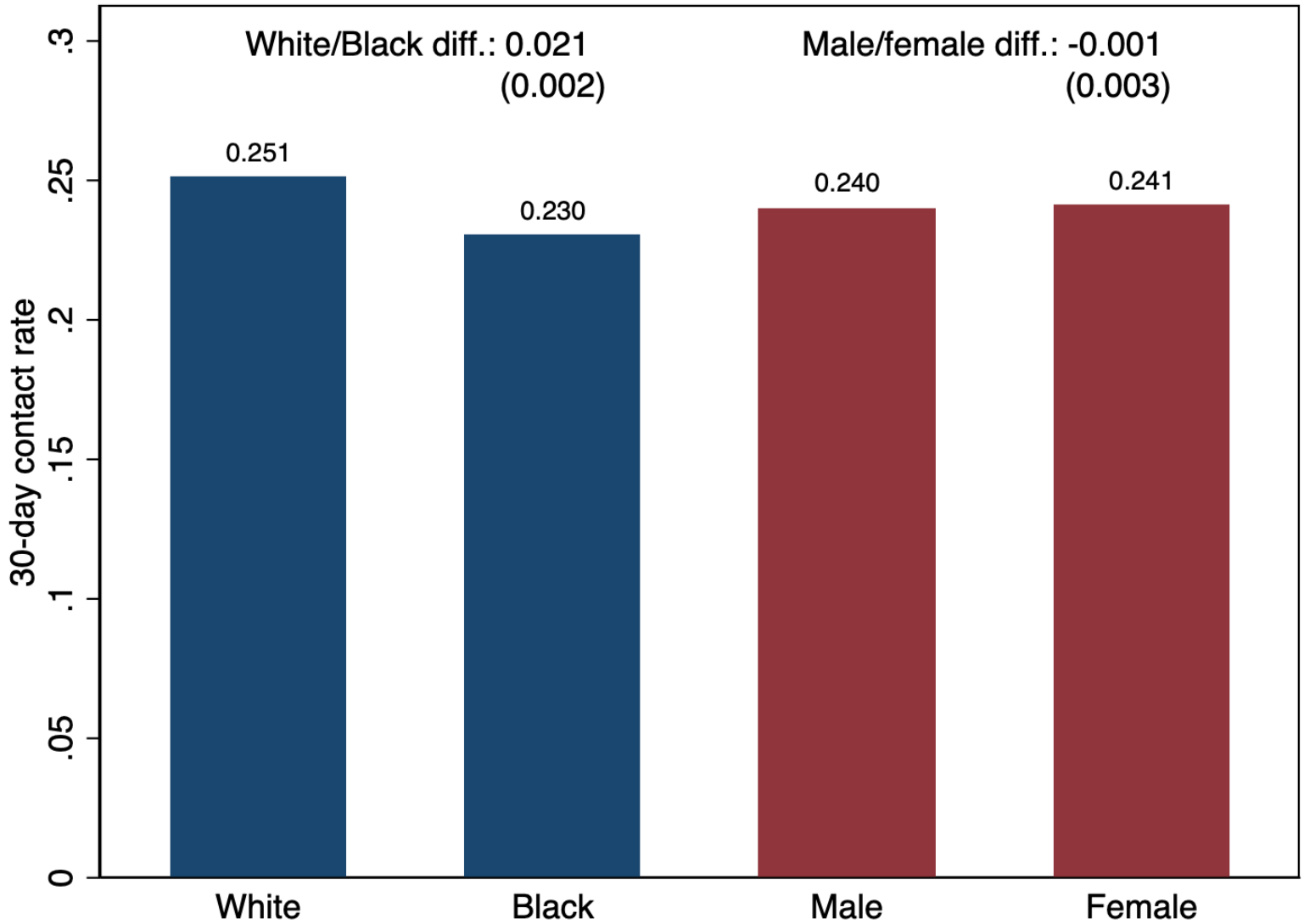
Skills

Communication
Prioritizing tasks
Highly detail oriented

Summary stats

	A. All firms			B. Balanced sample		
	White	Black	Difference	White	Black	Difference
Resume characteristics						
Female	0.499	0.499	-0.001	0.500	0.498	0.003
Over 40	0.535	0.535	0.000	0.534	0.533	0.002
LGBTQ club member	0.081	0.082	-0.001	0.079	0.080	-0.001
Academic club	0.040	0.042	-0.002	0.039	0.042	-0.003*
Political club	0.042	0.042	0.001	0.042	0.041	0.001
Gender-neutral pronouns	0.041	0.041	-0.001	0.040	0.040	0.000
Same-gender pronouns	0.043	0.042	0.001	0.042	0.041	0.001
Associate degree	0.476	0.485	-0.009**	0.478	0.485	-0.006*
N applications	41837	41806	83643	32703	32665	65368
N jobs			11114			8667
N firms			108			72
1/2/3/4/5 waves			3/4/15/16/72			

Average Contact Gaps by Race and Gender



Job-level Estimates

- Let $Y_{ijf}(r) \in \{0, 1\}$ indicate potential callback to applicant i at job j within firm f if assigned race $r \in \{b, w\}$
- Average treatment effect at this job is $\Delta_{jf} \equiv E[Y_{ijf}(w) - Y_{ijf}(b)]$
- Observed outcome is $Y_{ijf} = Y_{ijf}(R_{ijf})$, with $R_{ijf} \in \{b, w\}$
- White/Black difference in callback rates (**contact gap**):

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

- Random assignment of $R_{ijf} \implies \hat{\Delta}_{jf}$ is an unbiased estimate of Δ_{jf}

Firm-level Estimates

- Let $\theta_f = E_f[\Delta_{jf}]$ denote the average of Δ_{jf} across all jobs within firm f
- Observed average contact gap across the n_f jobs at firm f:

$$\hat{\theta}_f = \frac{1}{n_f} \sum_{j=1}^{n_f} \hat{\Delta}_{jf}$$

- Random sampling of jobs $\implies$ $\hat{\theta}_f$ is an unbiased estimate of θ_f
- Unbiased (squared) standard error:

$$s_f^2 = \frac{1}{n_f(n_f-1)} \sum_{j=1}^{J_f} (\hat{\Delta}_{jf} - \hat{\theta}_f)^2$$

- EB step 1: $\{\hat{\theta}_f, s_f\}_{f=1}^{F}$ provide building blocks for analysis of firm heterogeneity

The Variance of Discrimination

- Let G denote the distribution of contact gaps across firms: $\theta_f \sim G$
- Estimators for mean and variance of G:

$$\hat{\mu}_\theta = \frac{1}{F}\sum_{f=1}^{F} \hat{\theta}_f,$$

$$\hat{\sigma}^2_\theta = \frac{1}{F}\sum_{f=1}^{F} \left[\left(\hat{\theta}_f - \hat{\mu}_\theta \right)^2 - \left(\frac{F-1}{F} \right) s_f^2 \right]$$

- $\hat{\sigma}^2_\theta$ is a special case of Kline, Saggio and Sølvsten (2020) unbiased variance estimator
  - Unbiased s_f^2+ degrees of freedom correction $\implies$ finite-sample unbiased estimate
- We can then form linear shrinkage estimates $\hat{\theta}^*_f = \left(\frac{\hat{\sigma}^2_\theta}{\hat{\sigma}^2_\theta + s_f^2} \right) \hat{\theta}_f + \left(\frac{s_f^2}{\hat{\sigma}^2_\theta + s_f^2} \right) \hat{\mu}_\theta$

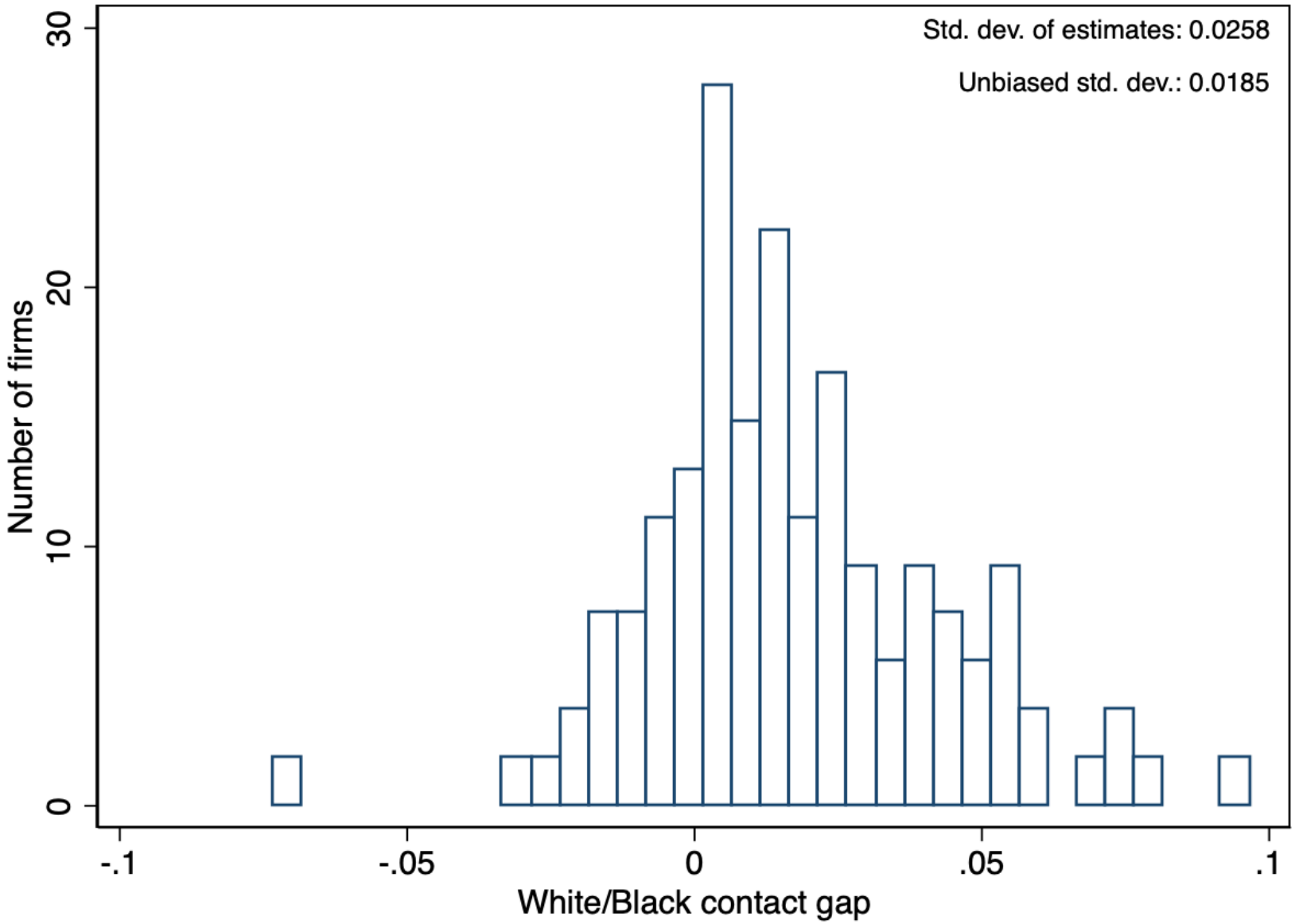
Standard Deviations of G: Substantial Variation for Both Race and Gender

Estimates of firm heterogeneity in race and gender discrimination

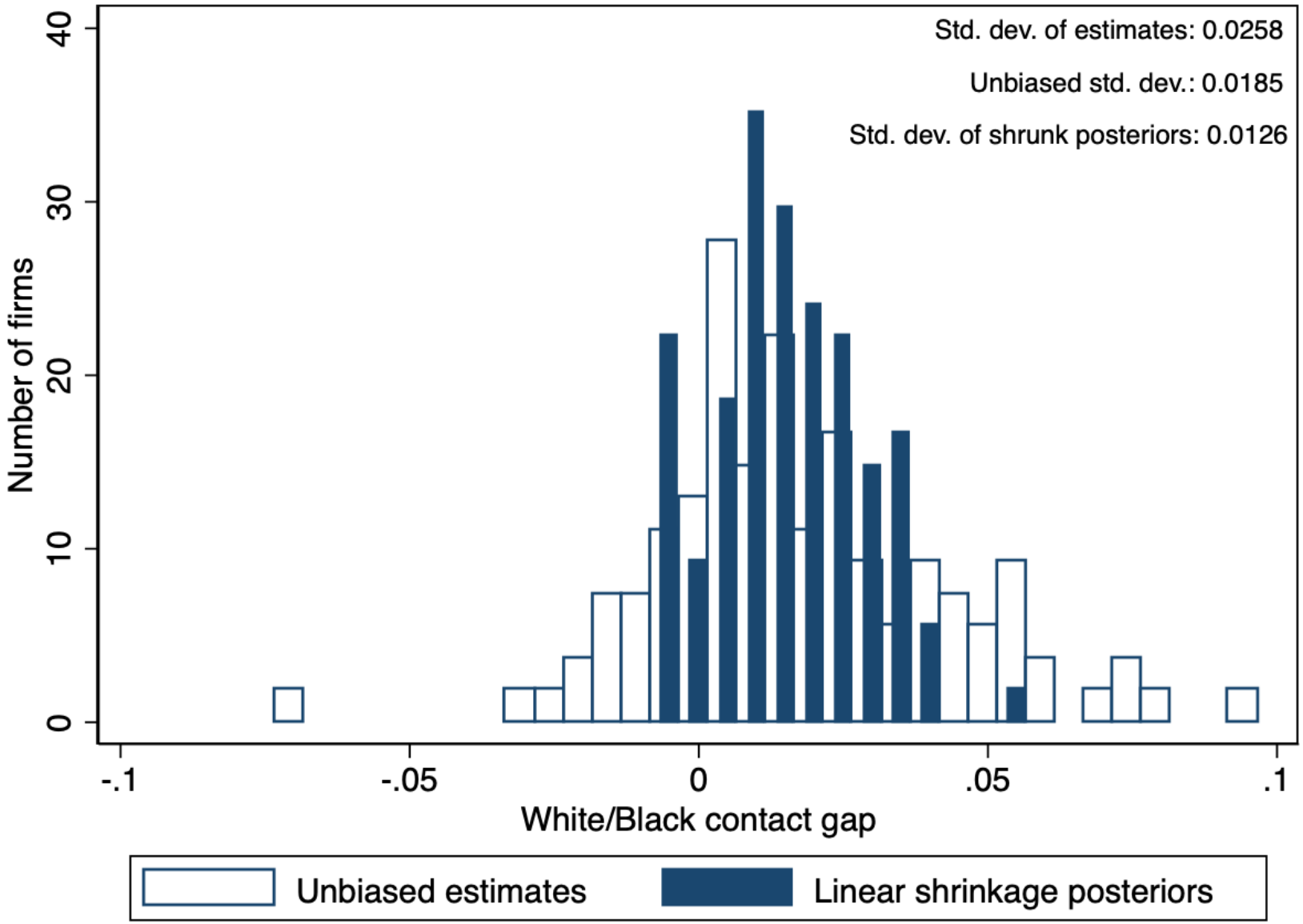
	Mean contact gap (1)	Bias-corrected std. dev. of contact gaps (2)
Race (White - Black)	0.021	0.0185
	(0.002)	(0.0031)
Gender (Male - Female)	-0.001	0.0267
	(0.003)	(0.0038)

Estimates from Kline, Rose, and Walters (2022).

Histogram of Unbiased Contact Gap Estimates



Unbiased vs. Linear Shrinkage Contact Gap Estimates



Non-Parametric Deconvolution

- To get a richer picture of G, return to hierarchical random effects framework with normal noise:

$$\hat{\theta}_j | \theta_j, s_j \sim N(\theta_j, s_j^2)$$

$$\theta_j \sim G$$

- As before, think of normality of $\hat{\theta}_j | \theta_j, s_j$ as an asymptotic approximation
- **Non-parametric deconvolution**: Estimate mixing distribution G under minimal assumptions on its shape/properties
- N.B.: Need to account for any dependence between effect sizes θ_j and sampling variances s_j^2
- Two approaches to non-parametric deconvolution:
  - Non-parametric maximum likelihood
  - Log-spline deconvolution

Non-Parametric Maximum Likelihood

- Classic deconvolution approach: **Non-parametric maximum likelihood estimator** (NPMLE; Robbins, 1950; Kiefer and Wolfowitz, 1956; Heckman and Singer, 1984)
- NPMLE picks mixing distribution to maximize likelihood of observed data:

$$\hat{G} = \arg\max_{G \in \mathcal{G}} \sum_{j=1}^{J} \log \left(\int \frac{1}{s_j} \phi \left(\frac{\hat{\theta}_j - \theta}{s_j} \right) dG(\theta) \right)$$

- Solution is a discrete distribution $\hat{G}$ with at most J mass points
- Koenker and Mizera (2014) develop an approximation that is straightforward to compute with modern convex optimization methods
  - Implemented in **REBayes** R package (Koenker and Gu, 2017)

Log-Spline Deconvolution

- Efron (2016) proposes to approximate G with distribution in smooth exponential family with log density parameterized by a natural spline
- For a set of M support points $(\bar{\theta}_1, ..., \bar{\theta}_M)$, suppose mass at point m is given by

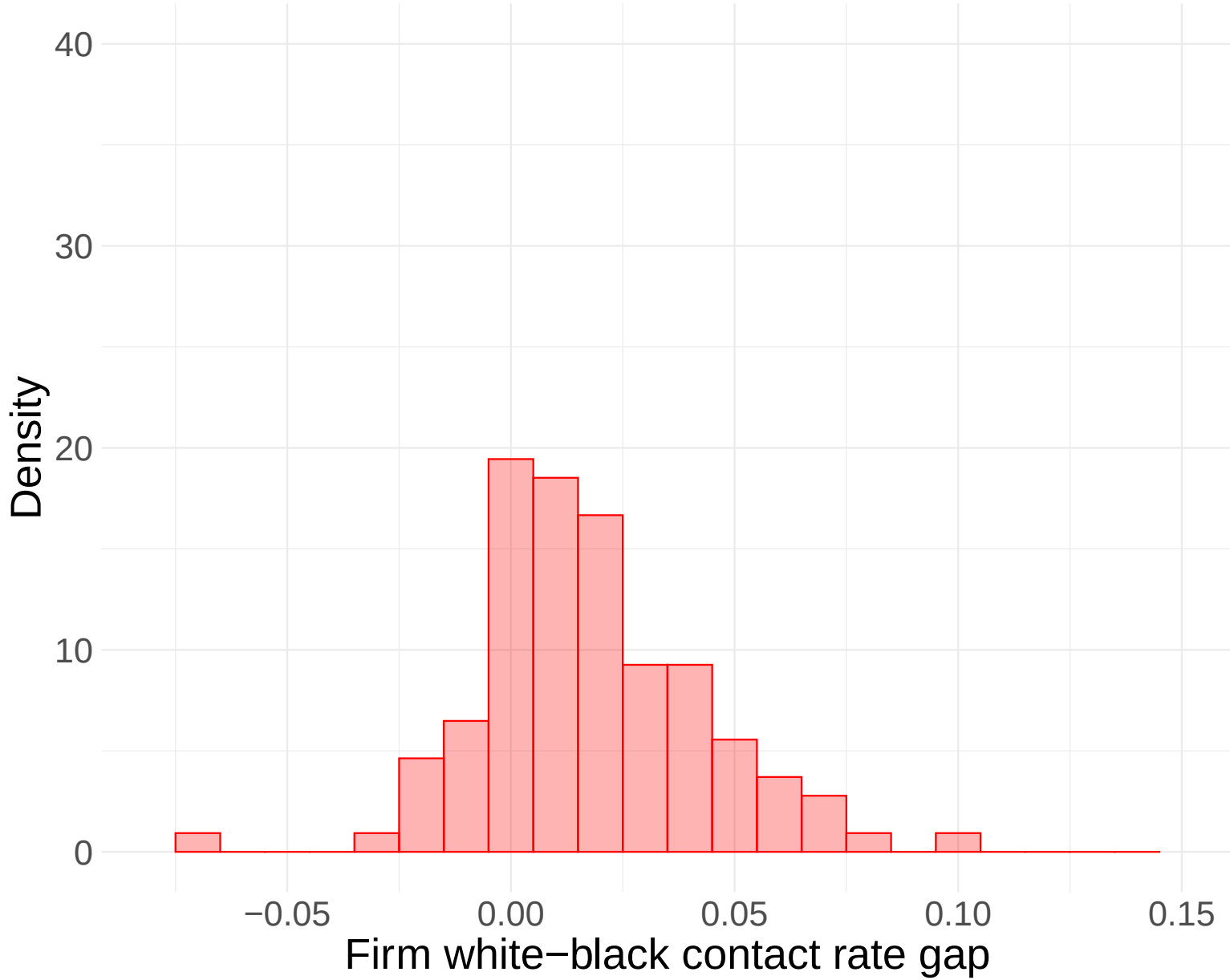
$$g_m(\alpha) = \exp\left(q_m'\alpha - \log\left(\sum_{\ell=1}^{M} \exp(q_\ell'\alpha)\right)\right)$$

- q_m is a $B \times 1$ vector of values of natural spline basis functions; α is a $B \times 1$ parameter vector
- Estimate α by penalized maximum likelihood:

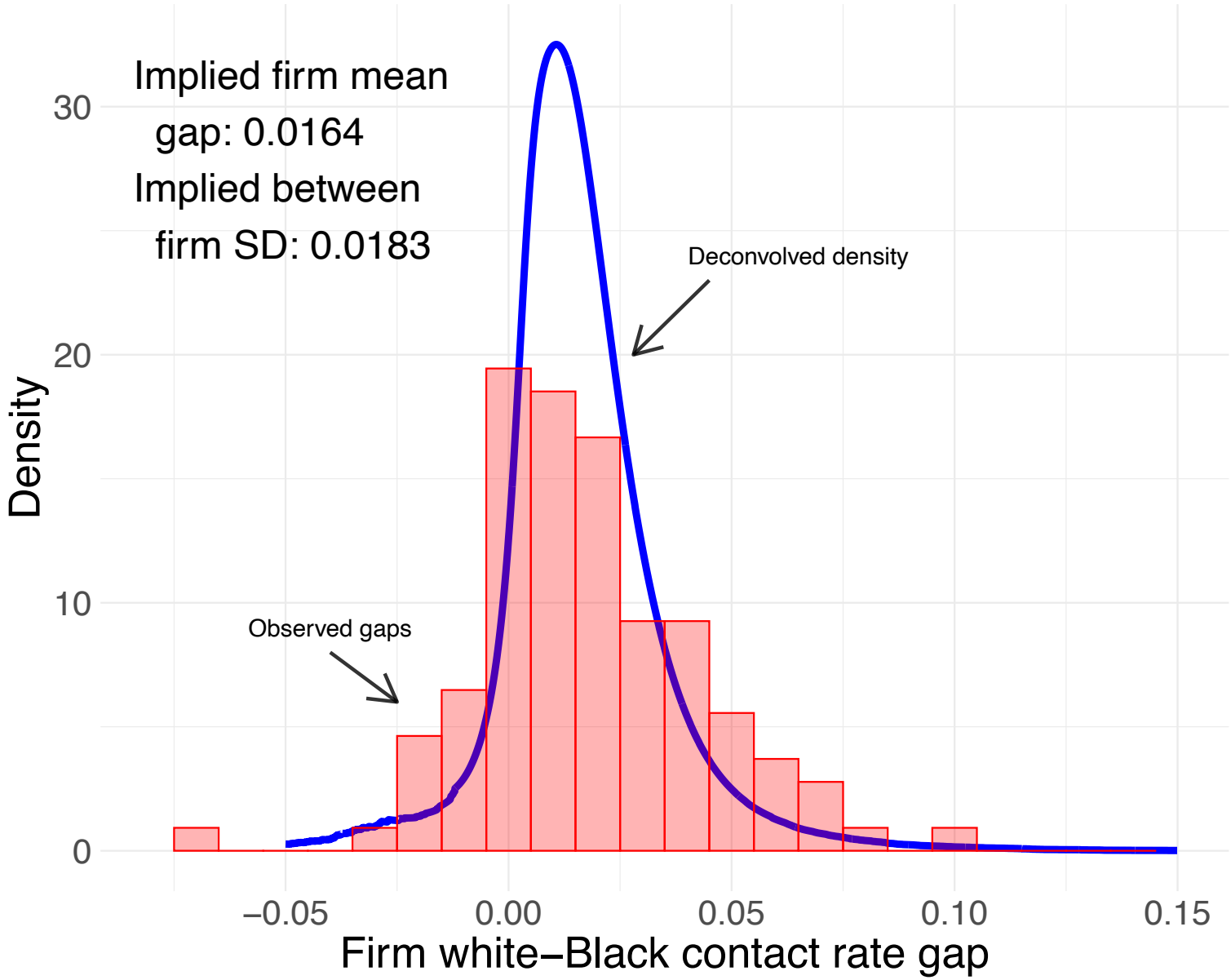
$$\hat{\alpha} = \arg\max_{\alpha} \sum_{j=1}^{J} \log\left(\sum_{m=1}^{M} g_m(\alpha)\frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j - \bar{\theta}_m}{s_j}\right)\right) - c\sqrt{\alpha'\alpha}$$

- Requires choosing tuning parameters: penalty c, number and range of support points $\bar{\theta}_m$
- Homoskedastic-noise version implemented in **deconvolveR** R package (Narasimhan and Efron, 2020)

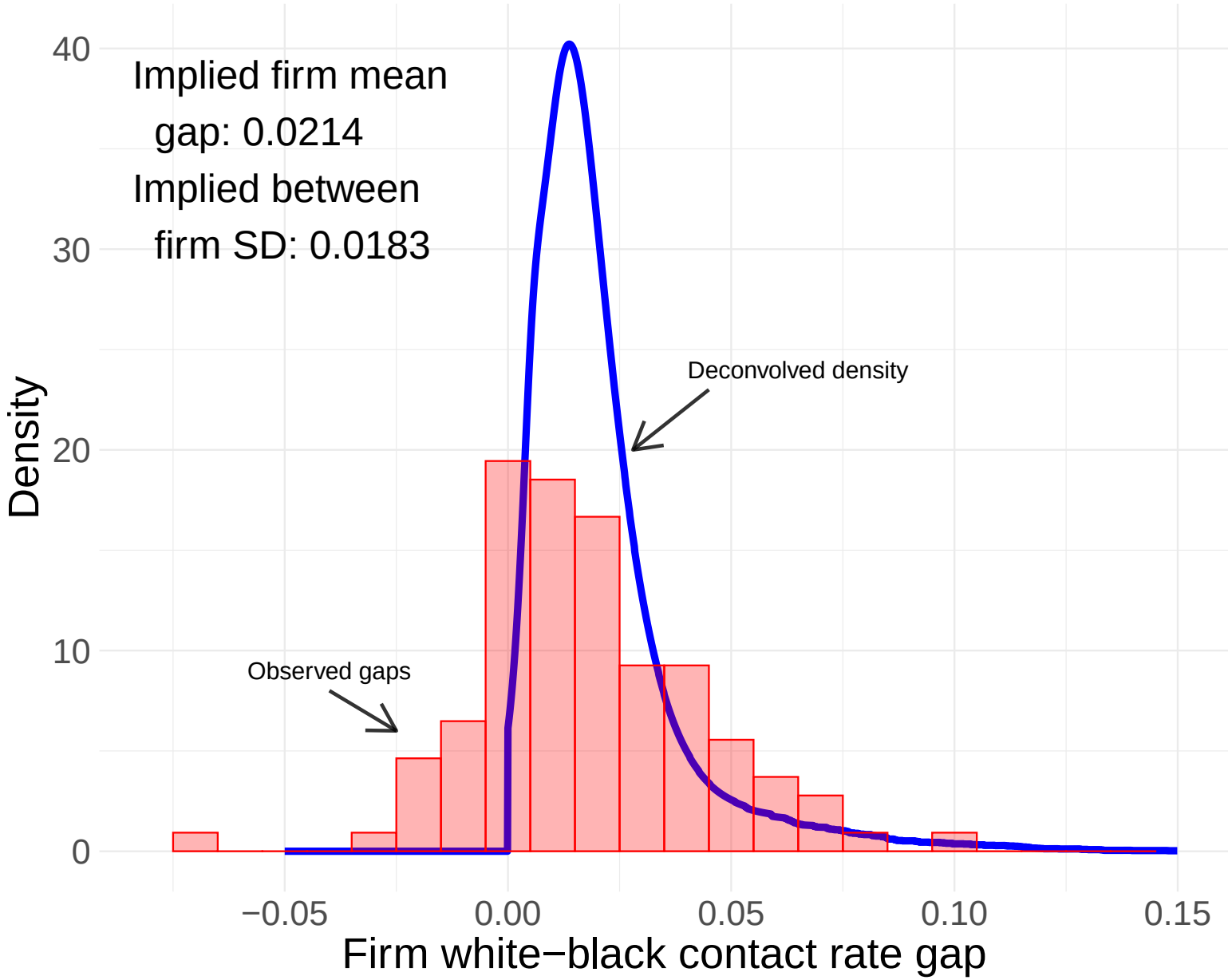
Histogram of Race Contact Gap Estimates



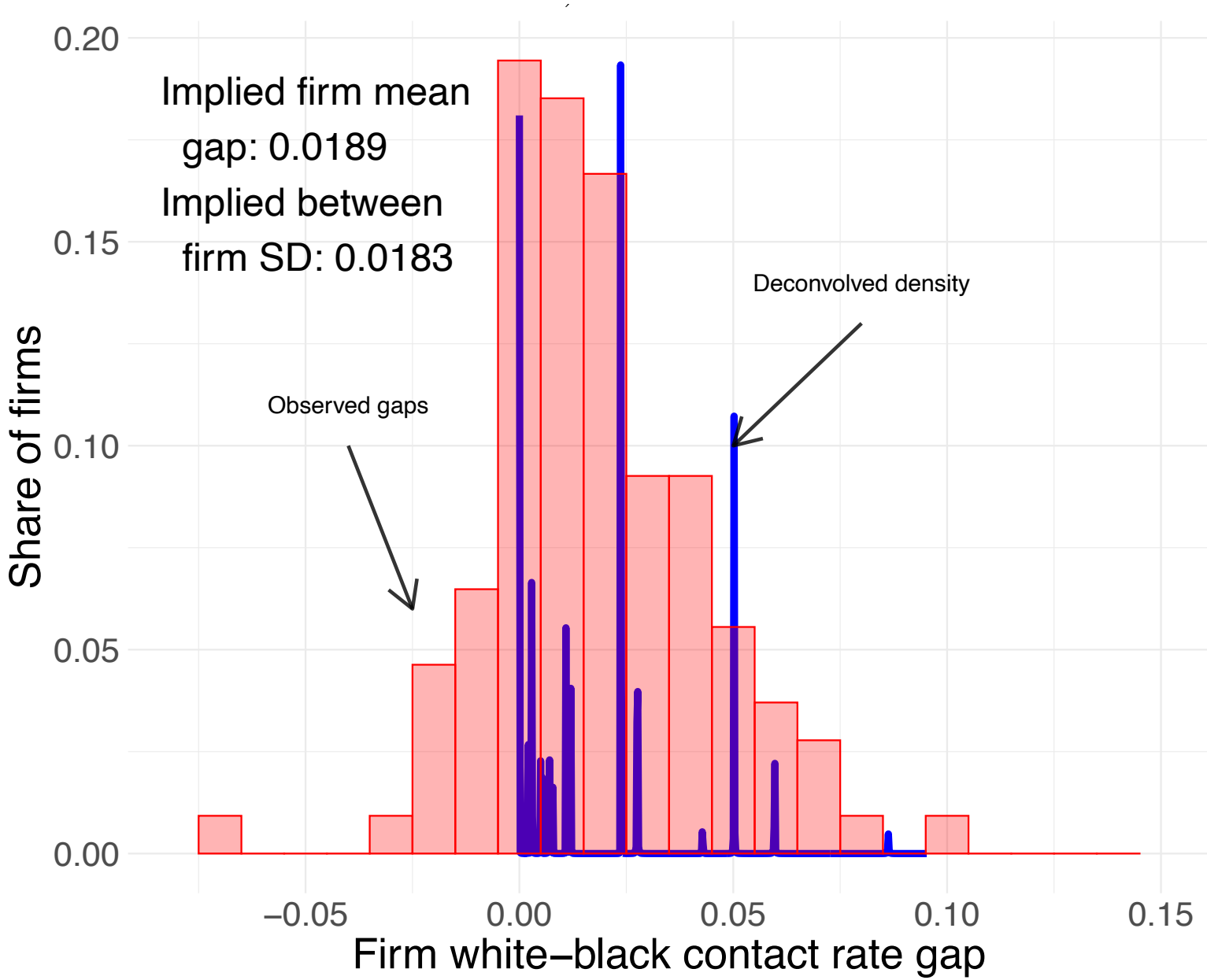
Log-spline Deconvolution Estimate for Race



Deconvolution Imposing Shape Restriction: $\theta_f \geq 0$

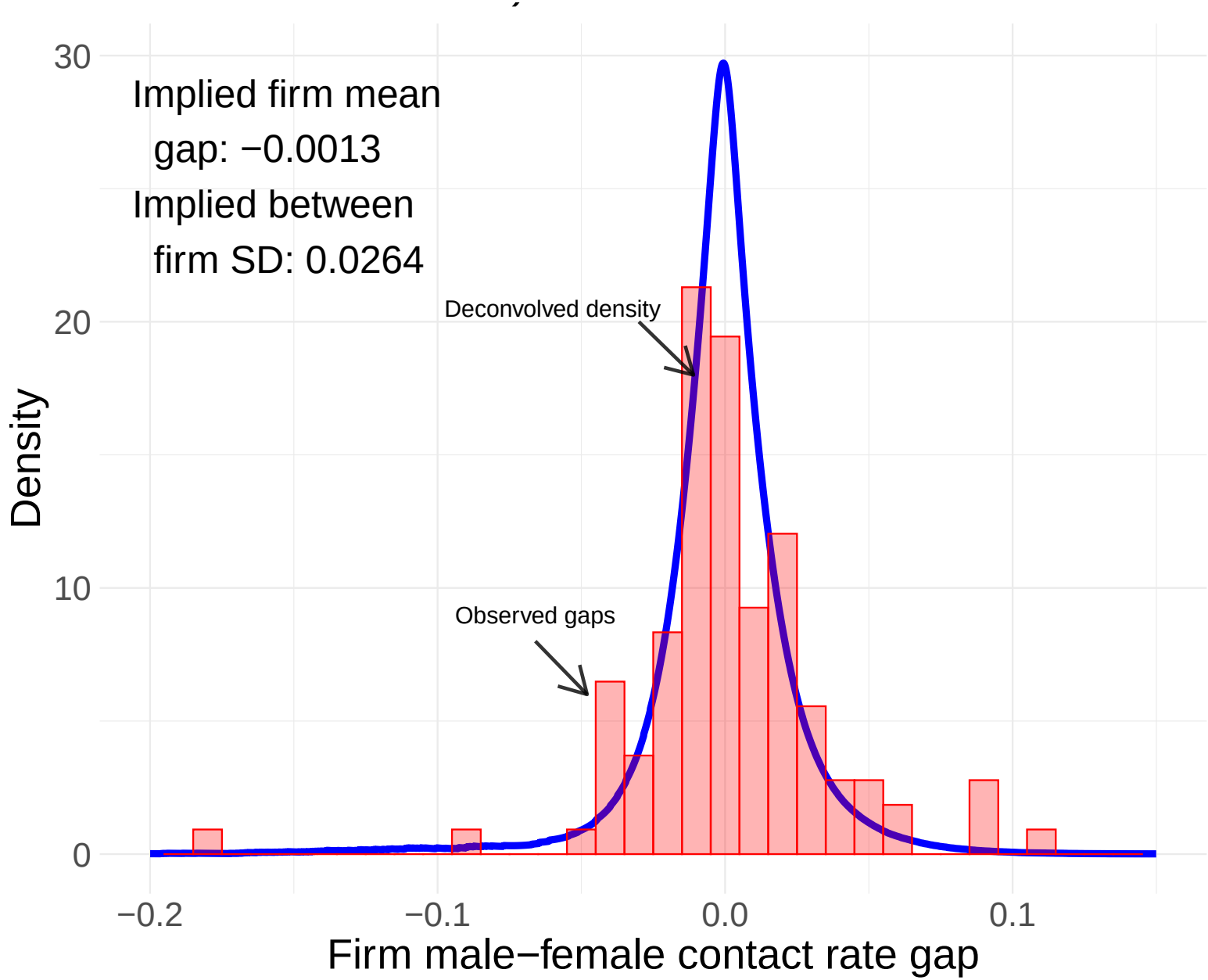


NPMLE Deconvolution Estimate for Race

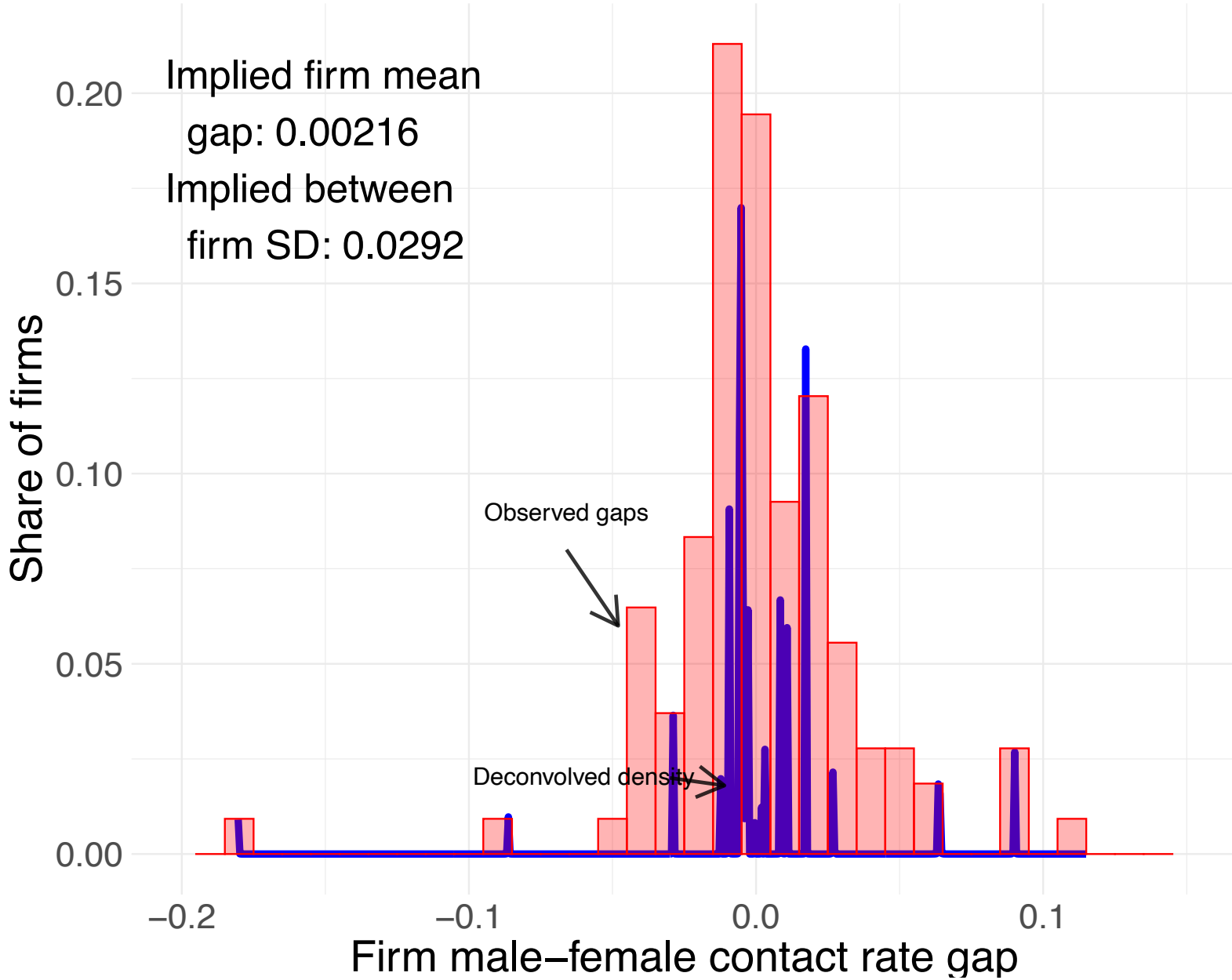


Histogram of Gender Contact Gap Estimates

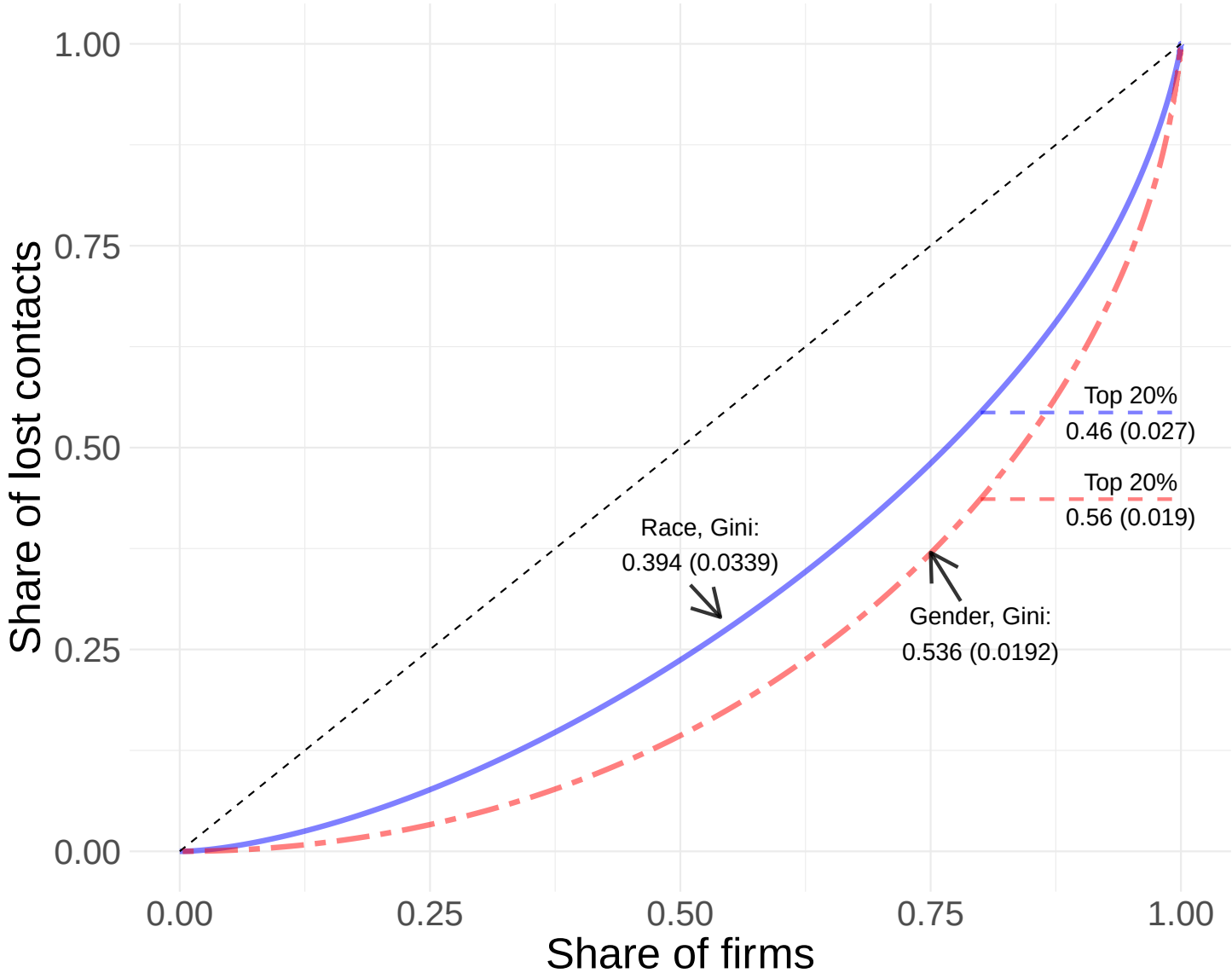
Log-spline Deconvolution Estimate for Gender



NPMLE Deconvolution Estimate for Gender



Lorenz Curves Derived from Log-spline $\hat{G}$'s



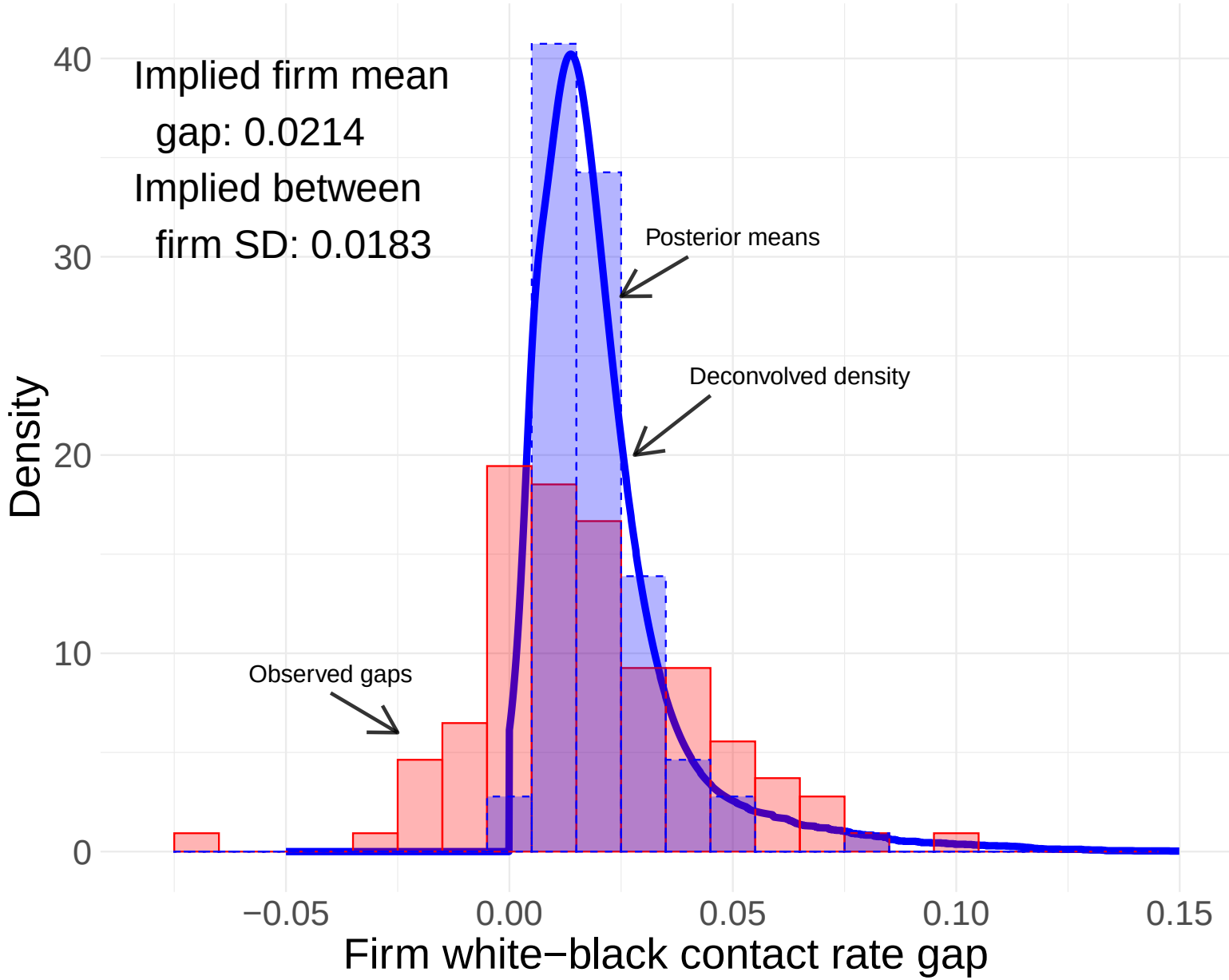
Non-Parametric Shrinkage

- After obtaining $\hat{G}$ via non-parametric deconvolution, we can form non-parametric posteriors using posterior density:

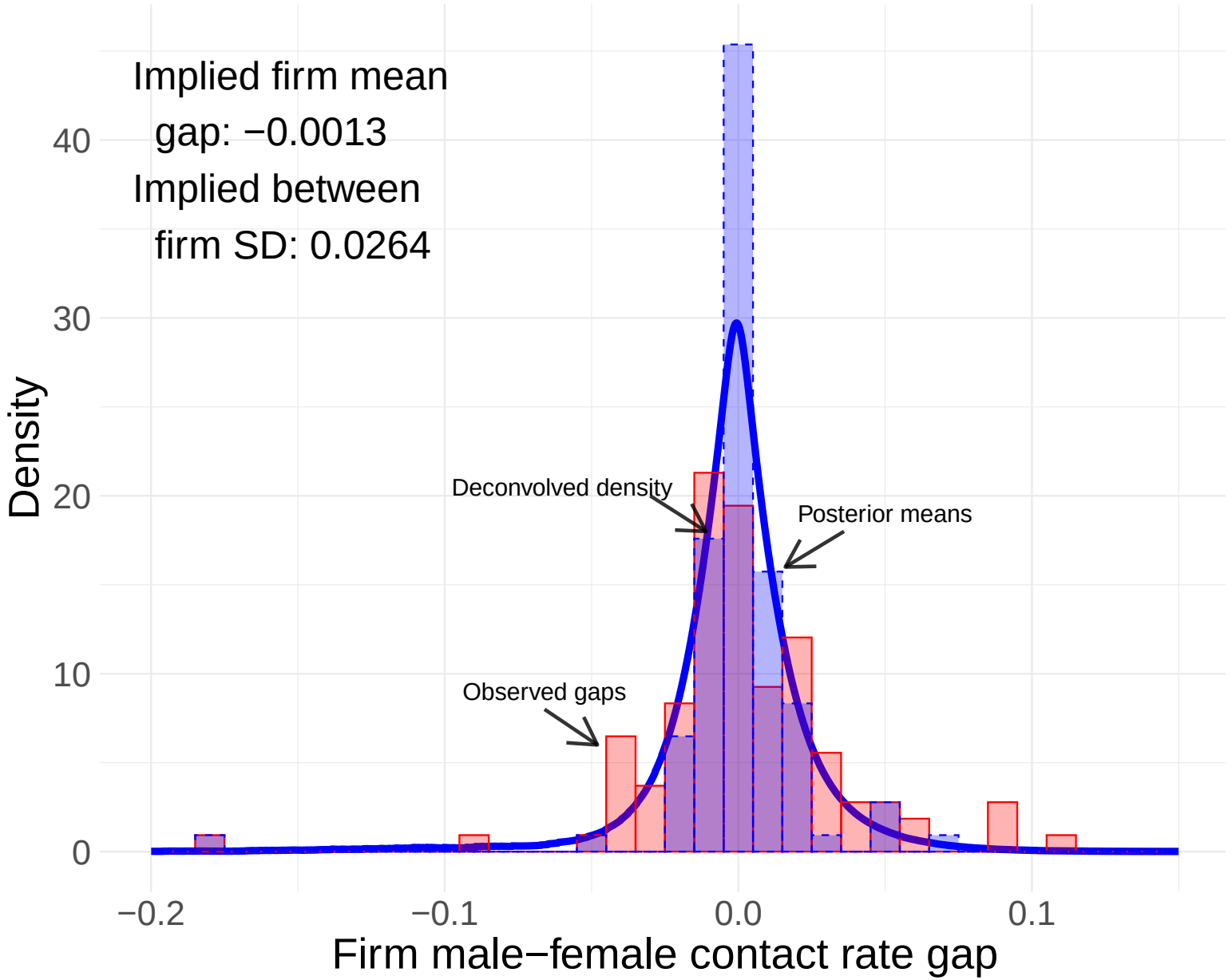
$$\hat{f}\left(\theta_j|\hat{\theta}_j, s_j\right) = \frac{\frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j-\theta_j}{s_j}\right) d\hat{G}\left(\theta_j\right)}{\int \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j-\theta}{s_j}\right) d\hat{G}(\theta)}$$

- Non-parametric posterior mean $\hat{\theta}_j^* = \int \theta \hat{f}(\theta|\hat{\theta}_j, s_j)d\theta$ will generally differ from linear shrinkage estimate $\tilde{\theta}_j = \left(\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2+s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\hat{\sigma}_\theta^2+s_j^2}\right)\hat{\mu}_\theta$
- $\hat{\theta}_j^*$ may be more accurate (lower MSE) if G is not normal
- On the other hand, linear shrinkage only requires estimating mean and variance of G
  - $\tilde{\theta}_j$is best linear approximation to true unknown posterior mean θ_j^*
  - May perform better than non-linear estimate $\hat{\theta}_j^*$ if higher moments of G are poorly-estimated

Histogram of Posterior Means for Race



Histogram of Posterior Means for Gender



Deconvolution with Precision-Dependence

- With precision-dependence, we need to estimate conditional mixing distribution $G(\theta|s_j)$
- One approach is to assume or estimate a model of the relationship between θ_j and s_j
- For example, we might allow dependence only through the mean:

$$E[\theta_j|s_j] = \alpha + \beta \log s_j,$$

$$r_j \equiv \theta_j - \alpha - \beta \log s_j,$$

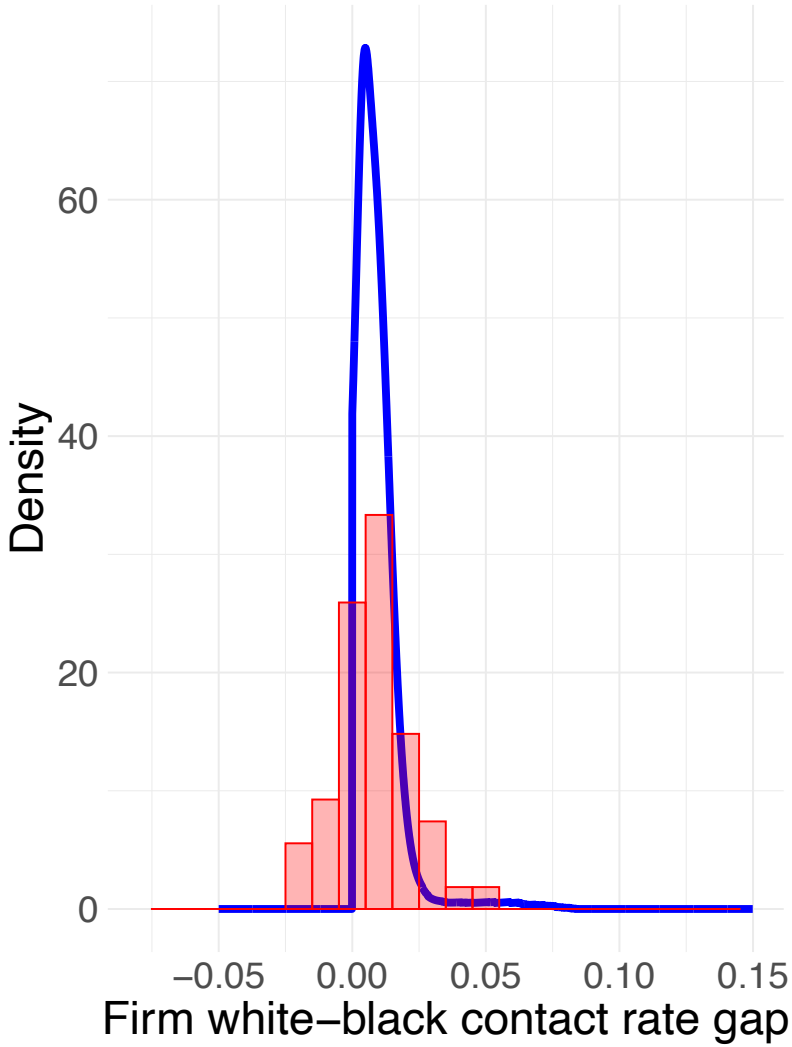
$$r_j|s_j \sim G_r$$

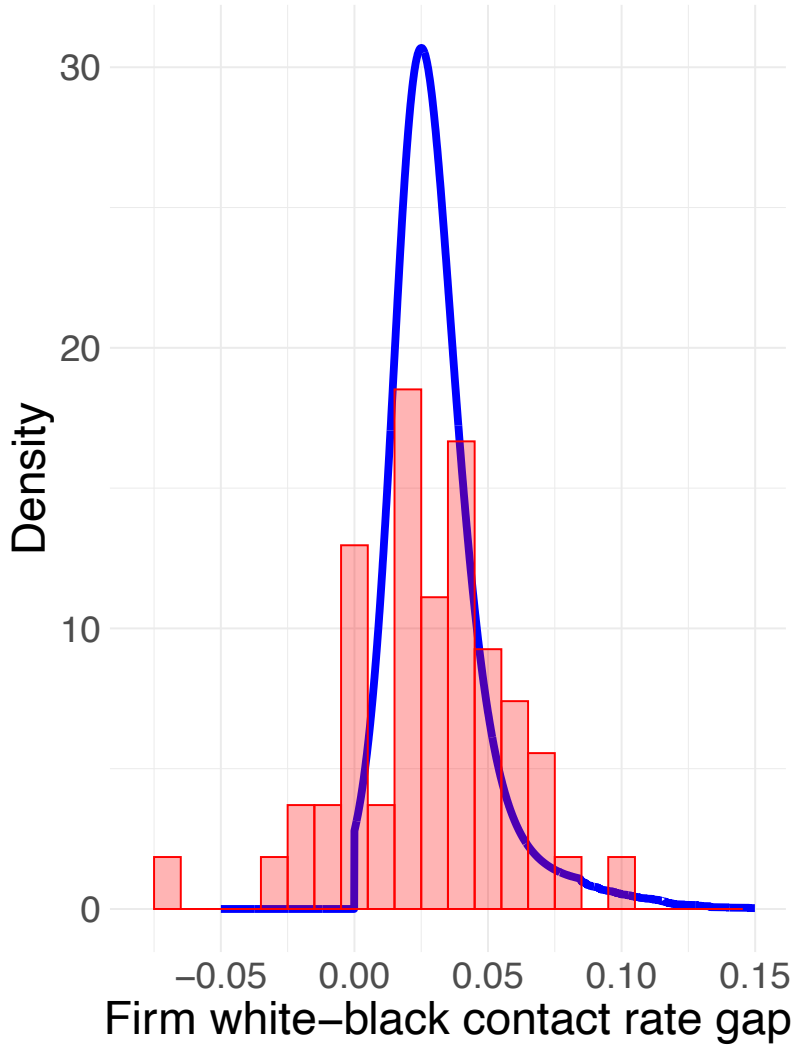
- Regress $\hat{\theta}_j$ on $\log s_j$, then deconvolve residual $\hat{r}_j$ to estimate G_r
- Can generalize to location/scale models or more complicated models of dependence (Chen, 2023)

Using z-score Transformations

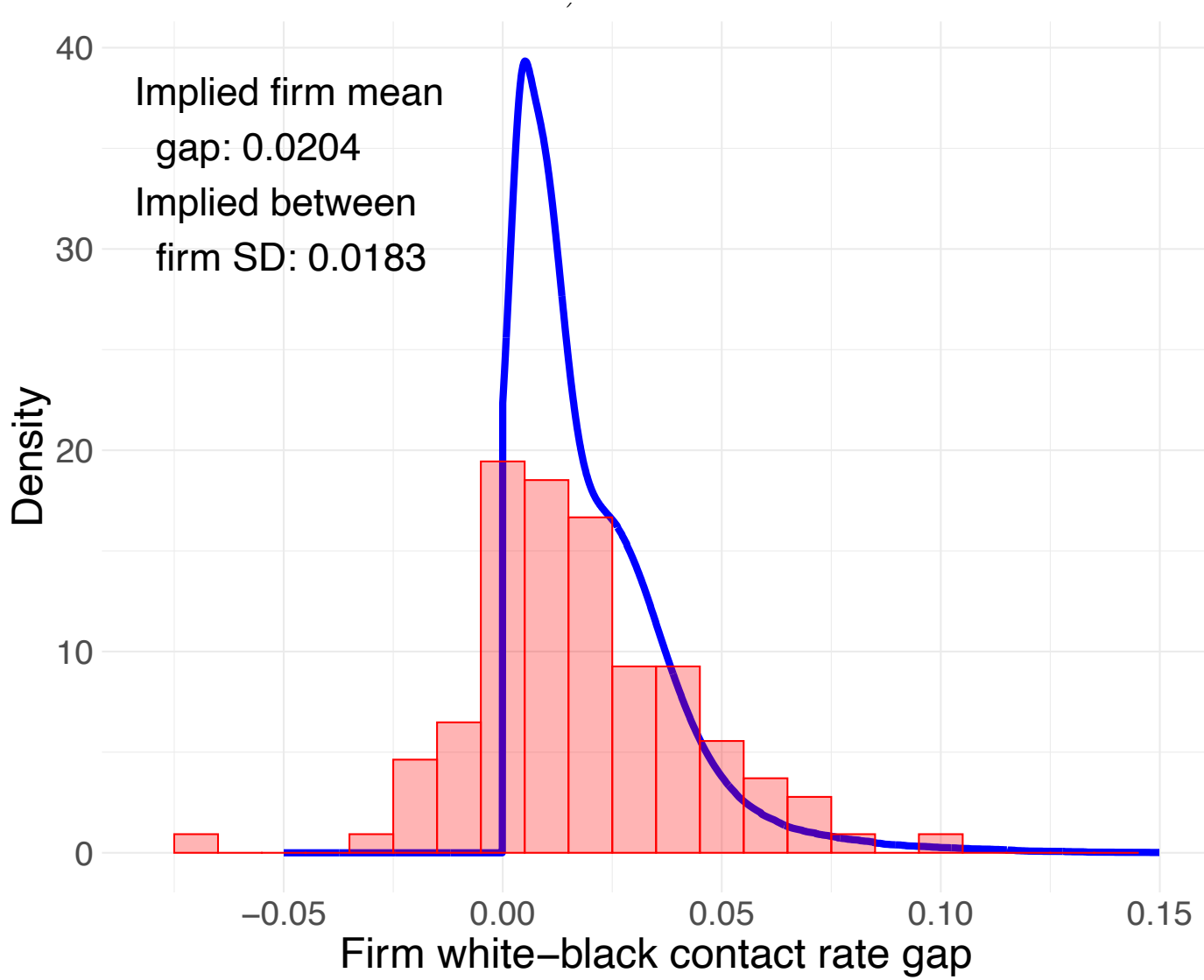
- Kline, Rose and Walters (2022) initially sidestep precision-dependence by transforming estimates to z-scores $z_f = \hat{\theta}_f / s_f$
  - $z_f \sim N(\mu_f, 1)$, where $\mu_f = \theta_f / s_f$
  - Deconvolve z-scores to estimate distribution of μ_f
- Then reconstruct distribution of $\theta_f = s_f \mu_f$ by a change of variables
  - If μ_f is independent of s_f, then $g_\theta(\theta) = \int \frac{1}{s} g_\mu(\theta/s) h_s(s) ds$
- Note that independence of μ_f from s_f implies $E[\theta_f | s_f] = E[\mu_f] \times s_f$, which is increasing in s_f
  - Seems to provide good empirical fit
  - Alternative approach: deconvolve separately in bins of s_f; yields similar results

Separate Deconvolutions for Low vs. High s_f





Marginal Distribution from Separate Deconvolutions



Variance-stabilizing Transformations

- An alternative approach is to eliminate heteroskedasticity with a **variance-stabilizing transformation** (VST)
- Suppose $\hat{\theta}_j|\theta_j, s_j^2 \sim N(\theta_j, s_j^2)$
- In addition, suppose sampling variance s_j^2 is a known deterministic function of effect size θ_j:

$$s_j^2 = h(\theta_j)$$

- Define $t(\theta) \equiv v \int_{-\infty}^{u} h(u)^{-1/2} du$. By the delta method, we have

$$t(\hat{\theta}_j)|\theta_j \sim N(t(\theta_j), v^2)$$

- We can then deconvolve the transformed variable $t(\hat{\theta}_j)$, which has homoskedastic noise

Binomial VST

- Suppose we have binomial data with success probability θ_j in group j and N trials per group:

$$Y_j \sim Bin(N, \theta_j)$$

- The sample proportion of successes $\hat{\theta}_j = Y_j/N$ has $E[\hat{\theta}_j|\theta_j] = \theta_j$ and $Var(\hat{\theta}_j|\theta_j) = \theta_j(1-\theta_j)/N$
- Classic VST for a binomial (Bartlett, 1936, 1947; Anscomb, 1948):

$$t(\theta_j) = \arcsin\sqrt{\theta_j}$$

- For the binomial, $Var(t(\hat{\theta}_j)|\theta_j) = \frac{1}{4N} + O(N^{-2})$, which no longer depends on θ_j
- Brown (2008) considers a wider class of VSTs for binomial models

VST Drawbacks

- With more complicated data structures a known VST will typically not be available
- Can try to estimate a VST by fitting a model of the form $s_j^2 = h(\theta_j; \gamma)$ for unknown parameters γ
  - But then we are essentially back to modeling the dependence
- Outside of special cases it is also not clear when we should expect a deterministic relationship between precision and effect sizes

Noise in Estimates of Precision

- A related problem is that estimated standard errors s_j may be noisy estimates of sampling uncertainty in $\hat{\theta}_j$'s
- Consider classic normal means problem with $Y_{ij} \sim N(\theta_j, \sigma_j^2)$
- Estimators of the mean and variance: $\hat{\theta}_j = N_j^{-1} \sum_i Y_{ij}$, $\hat{\sigma}_j^2 = (N_j - 1)^{-1} \sum_i (Y_{ij} - \hat{\theta}_j)^2$
- Distribution of estimated mean and variance:

$$\hat{\theta}_j \sim N\left(\theta_j, \sigma_j^2 / N_j\right), \; (N_j - 1)\hat{\sigma}_j^2 / \sigma_j^2 \sim \chi^2_{N_j - 1},$$

$$\hat{\theta}_j \perp\!\!\!\perp \hat{\sigma}_j^2$$

- This implies squared SE $\hat{s}_j^2 = \hat{\sigma}_j^2 / N_j$ follows a Gamma distribution with shape $(N_j - 1)/2$ and scale $2\sigma_j^2 / [N_j(N_j - 1)]$, independent of $\hat{\theta}_j$

Dealing with Noisy Standard Errors

- In more general settings we may have dependence between noise in $\hat{\theta}_j$ and $\hat{s}_j$, in addition to dependence between θ_j and s_j
- Bivariate deconvolution: jointly deconvolve $(\hat{\theta}_j, \hat{s}_j)$ to recover bivariate mixing distribution $G(\theta, s)$
  - NPMLE version implemented in **GLVmix** R package of Koenker and Gu (2017)
- Could try multivariate VST (covariance-stabilizing transform), though this does not always exist (Holland, 1973)
- In the normal means case, $Var(\hat{s}_j^2) = 2\sigma_j^4/[N_j^2(N_j - 1)] = 2s_j^4/(N_j - 1)$
  - Goes to zero with N_j faster than $s_j^2 = \sigma_j^2/N_j$
  - Perhaps for this reason, precision is typically treated as known in practice

Large-Scale Inference

- Empirical Bayes methods are closely related to multiple testing approaches ("large-scale inference;" Efron, 2012)
- Suppose we are interested in a null hypothesis involving θ_j for each unit, e.g., $H_0 : \theta_j = 0$
  - Which subgroups are affected by an intervention?
  - Which firms discriminate against Black applicants?
- Let $T_j \in \{0, 1\}$ denote an indicator equal to1 if the null is true for unit j
- Let $R_j \in \{0, 1\}$ denote an indicator equal to 1 if we reject the null for unit j

Multiple Testing

- Traditional hypothesis testing controls the probability of type I error (size) for a single test:
  - Limit probability of a false rejection assuming the null is true
  - In other words, adopt a rejection rule such that $\Pr[R_j = 1 | T_j = 1] \leq \alpha$
- When we are interested in many hypotheses simultaneously, it is no longer clear what notion of error we should control
  - Family-wise error rate (FWER): Probability of at least one mistaken rejection
  - False Discovery Rate (FDR): Expected share of rejections that are mistaken
- For large-scale testing problems FWER control can be very stringent
- We will focus on controlling FDR, which is natural in the EB framework

The False Discovery Rate

- Suppose our tests yield p-values $p_1, ..., p_J$, and we reject those with $p_j \leq \bar{p}$. How many mistakes do we expect to make?
- By Bayes' rule, the share of true nulls among hypotheses we've rejected is:

$$\Pr[T_j = 1 | p_j \leq \bar{p}] = \frac{\Pr[p_j \leq \bar{p} | T_j = 1] \Pr[T_j = 1]}{\Pr[p_j \leq \bar{p}]}$$

$$= \frac{\bar{p}\pi_0}{F_p(\bar{p})}$$

- This quantity is the **False Discovery Rate** (FDR) for our decision rule (Benjamini and Hochberg, 1995)
- FDR is the expected share of true nulls among the hypotheses we reject
- If we can limit FDR, we can be reasonably confident that rejected hypotheses are false

FDR and G

$$FDR(\bar{p}) = \frac{\bar{p}\pi_0}{F_p(\bar{p})}$$

- P-values are uniformly distributed under the null, so $\Pr[p_j \leq \bar{p} | T_j = 1] = \bar{p}$
- The denominator is the marginal CDF of p-values, estimable from empirical share below $\bar{p}$
- Key unknown quantity is $\pi_0 = \Pr[T_j = 1]$, the share of true nulls in the population
  - If $\pi_0 = 1$, all rejected hypotheses are true
  - If $\pi_0 = 0$, all rejected hypotheses are false, but so are all hypotheses we don't reject
- π_0 is a feature of G: $\pi_0 = \int 1[\theta = 0] dG(\theta)$
- Can we use EB methods to get traction on π_0?

Bounding π_0

$$FDR(\bar{p}) = \frac{\bar{p}\pi_0}{F_p(\bar{p})}$$

- Conservative approach: plug in $\pi_0 = 1$ (Benjamini and Hochberg, 1995)
  - Still implies low FDR if many p-values close to 0 ($F_p(\bar{p}) >> \bar{p}$)
- But we can do better
  - Logically inconsistent to have $\pi_0 = 1$ but $F_p(\bar{p}) >> \bar{p}$
  - π_0 can't be 1 if mean or variance of $G \neq 0$
  - We can borrow strength from the ensemble of tests to bound π_0

Bounding π_0

- At any point p, density of p-values is mixure of true nulls (uniform) and false nulls (something else):

$$f_p(p) = \pi_0 + (1 - \pi_0) f_1(p)$$

- Since $f_1(p) \geq 0$, we have $\pi_0 \leq f_p(p)$ for any p, so minimum density of p-values bounds π_0 (Efron et al., 2001):

$$\pi_0 \leq \min_p f_p(p)$$

- Simple approach (Storey, 2002): calculate $\hat{\pi}_0$ as average density of p-values above threshold λ, beyond which we expect few false nulls
  - Formally, $\hat{\pi}_0 = \dfrac{\sum_{j=1}^{J} 1\{p_j > \lambda\} p_j}{(1 - \lambda) J}$
  - Higher λ means tighter bound but noisier estimate; Storey et al. (2004) discuss approaches to choosing λ

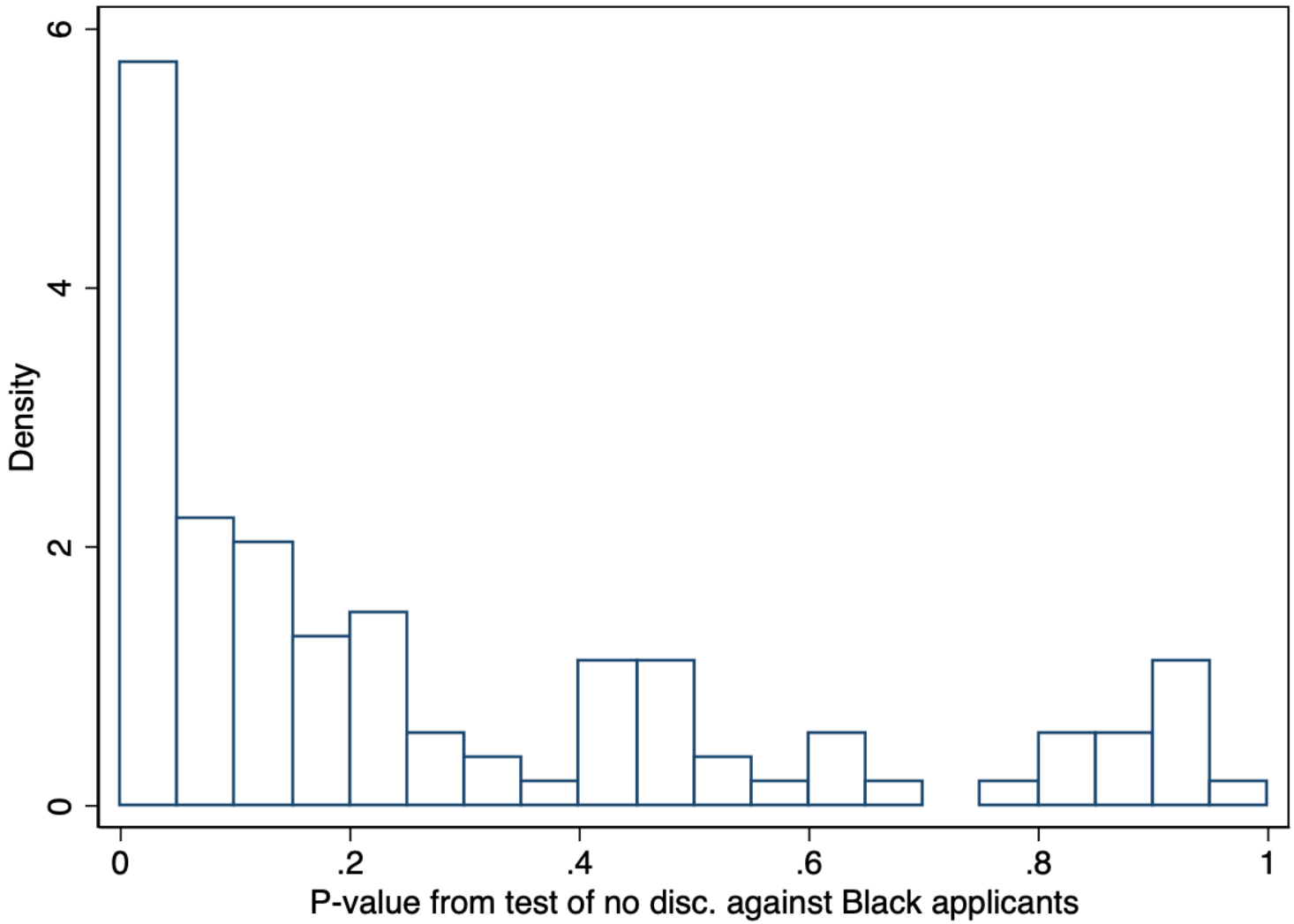
q-values for FDR Control

- Given estimated bound $\hat{\pi}_0$, control FDR using **q-values** (Storey, 2003):

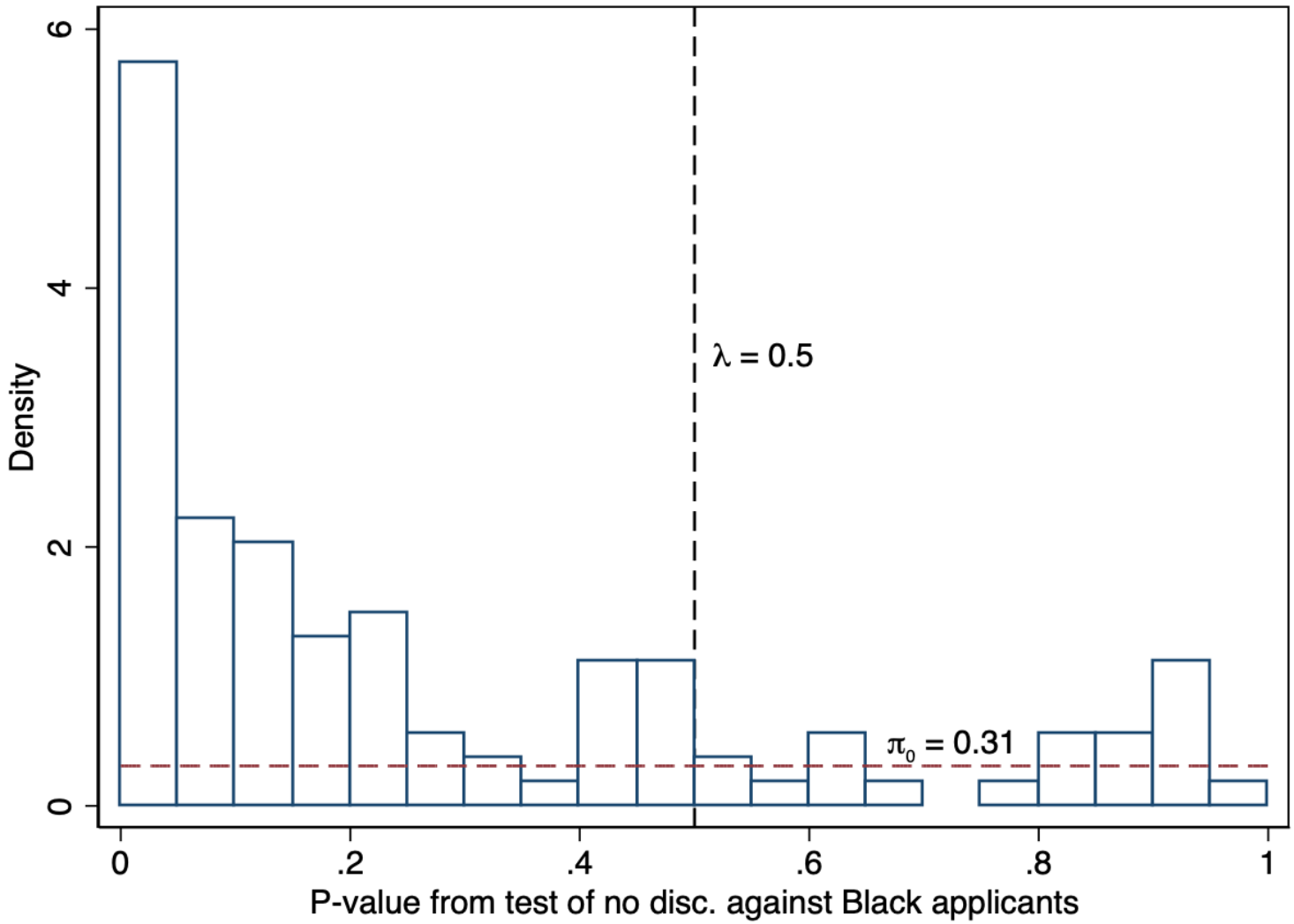
$$q_j = \widehat{FDR}(p_j) = \frac{p_j \hat{\pi}_0}{\hat{F}_p(p_j)}$$

- q-value $\approx$ EB equivalent of p-value
  - Rather than controlling $\Pr[R_j = 1 | T_j = 1]$, use Bayes rule + ensemble of tests to flip the conditioning and control $\Pr[T_j = 1 | R_j = 1]$
- If unit j's q-val is q_j and we reject all hypotheses with p-vals lower than p_j, we should expect at most $100q_j\%$ of rejections to be mistakes

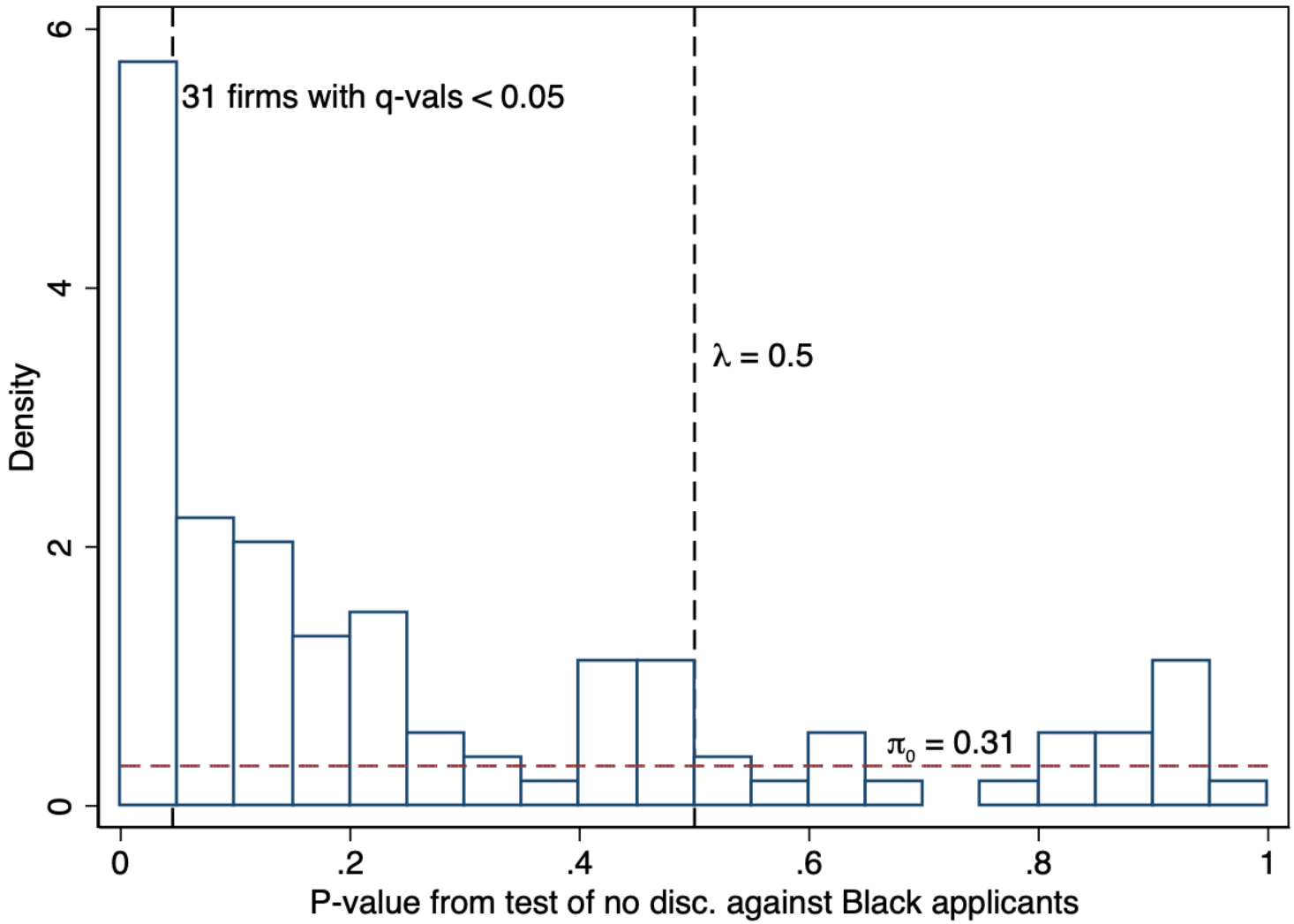
P-values from One-tailed Tests of No Discrimination



$\hat{\pi}_0 = 0.31 \implies$ At Least 69% of Firms Discriminate



31 Firms with q-vals Below 0.05



Industries and q-values from KRW (2022)

Firm	Industry	Contact gap estimate	Std. err.	p-value	q-value	Posterior mean
1	Auto dealers/services	0.0952	0.0197	0.0000	0.0001	0.0835
2	Auto dealers/services	0.0507	0.0143	0.0003	0.0061	0.0354
3	Auto dealers/services	0.0738	0.0220	0.0005	0.0073	0.0489
4	Auto dealers/services	0.0787	0.0249	0.0010	0.0103	0.0498
5	Apparel stores	0.0733	0.0250	0.0022	0.0158	0.0448
6	Other retail	0.0469	0.0159	0.0020	0.0158	0.0286
7	Other retail	0.0605	0.0219	0.0033	0.0176	0.0365
8	General merchandise	0.0520	0.0187	0.0031	0.0176	0.0314
9	Auto dealers/services	0.0613	0.0240	0.0060	0.0194	0.0370
10	Other retail	0.0560	0.0214	0.0050	0.0194	0.0337
11	Eating/drinking	0.0560	0.0222	0.0064	0.0194	0.0339
12	Auto dealers/services	0.0540	0.0215	0.0068	0.0194	0.0327
13	Food stores	0.0511	0.0204	0.0069	0.0194	0.0310
14	General merchandise	0.0427	0.0170	0.0068	0.0194	0.0259
15	Furnishing stores	0.0400	0.0159	0.0066	0.0194	0.0242
16	Wholesale nondurable	0.0386	0.0158	0.0080	0.0199	0.0235
17	Apparel manufacturing	0.0350	0.0142	0.0078	0.0199	0.0213
18	Building materials	0.0373	0.0157	0.0093	0.0218	0.0229
19	Health services	0.0544	0.0240	0.0132	0.0292	0.0339
20	Furnishing stores	0.0400	0.0183	0.0152	0.0322	0.0252
21	Eating/drinking	0.0340	0.0159	0.0172	0.0346	0.0217
22	General merchandise	0.0423	0.0210	0.0229	0.0439	0.0277
23	Insurance/real estate	0.0278	0.0140	0.0257	0.0472	0.0183

References

- Abadie, A., and Kasy, M. (2019). "Choosing among regularized estimators in empirical economics: the risk of machine learning." *Review of Economics and Statistics* 101(5).
- Abowd, J., Kramarz, F., and Margolis, D. (1999). "High-wage workers and high-wage firms." *Econometrica* 67(2).
- Angrist, J., Hull, P., Pathak, P., and Walters, C. (2017). "Leveraging lotteries for school value-added: testing and estimation." *Quarterly Journal of Economics* 132(2).
- Angrist, J., Hull, P., Pathak, P., and Walters, C. (forthcoming). "Simple and credible value-added estimation using centralized school assignment." *Review of Economics and Statistics.*
- Anscombe, F. (1948). "The transformation of poisson, binomial, and negative-binomial data." *Biometrika* 35(3/4).
- Bartlett, M. (1947). "The use of transformations." *Biometrics* 3(1).
- Benjamini, Y., and Hochberg, Y. (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 57(1).
- Bertrand, M., and Mullainathan, S. (2004). "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review* 94(4).

References

- Brown, L. (2008). "In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies." *The Annals of Applied Statistics* 2(1).
- Card, D. (1999). "The causal effect of education on earnings." *Handbook of Labor Economics* Volume 3.
- Card., D., Cardoso, A., Heining, J., and Kline, P. (2018). "Firms and labor market inequality: evidence and some theory." *Journal of Labor Economics* 36(S1).
- Chan, D., Gentzkow, M., and Yu, C. (2022). "Selection with variation in diagnostic skill: evidence from radiologists." *Quarterly Journal of Economics* 137(2).
- Chen, J. (2023). "Empirical Bayes when estimation precision predicts parameters." Working paper.
- Chetty, R., and Hendren, N. (2018). "The impacts of neighborhoods on intergenerational mobility II: county-level estimates." *Quarterly Journal of Economics* 133(3).Chetty, R., Friedman, J., Hendren, N., Jones, M., and Porter, S. (2018). "The opportunity atlas: mapping the childhood roots of social mobility." NBER working paper no. 25147.
- Chetty, R., Friedman, J., and Rockoff, J. (2014). "Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates." *American Economic Review* 104(9).
- Dale, S., and Krueger, A. (2002). "Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables." *Quarterly Journal of Economics* 117(4).

References

- Dale, S., and Krueger, A. (2014). "Estimating the effects of college characteristics over the career using administrative earnings data." *Journal of Human Resources* 49(2).
- Efron, B., and Morris, C. (1973). "Stein's estimation rule and its competitors – an empirical Bayes approach." *Journal of the American Statistical Association* 68(341).
- Einav., L., Finkelstein, A., and Mahoney, N. (2022). "Producing health: measuring value added of nursing homes." NBER working paper no. 30228.
- Efron, B. (2010). "The future of indirect evidence." *Statistical Science* 25(2).
- Efron, B. (2012). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Cambridge University Press.
- Efron, B. (2016). "Empirical Bayes deconvolution estimates." *Biometrika* 103(1).
- Efron, B., Tibshirani, R., Storey, J., and Tusher, V. (2001). "Empirical Bayes analysis of a microarray experiment." *Journal of the American Statistical Association* 96(456).
- Fenizia, A. (2022). "Managers and productivity in the public sector." *Econometrica* 90(3).

References

- Frandsen, B., Lefgren, L., and Leslie, E. (2023). "Judging judge fixed effects." *American Economic Review* 113(1).
- Goncalves, F., and Mello, S. (2021). "A few bad apples? Racial bias in policing." *American Economic Review* 111(5).
- Gu, J., and Koenker, R. (2023). "Invidious comparisons: ranking and selection as compound decisions." *Econometrica* 91(1).
- Heckman, J., and Singer, B. (1984). "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica* 52(2).
- Holland, P. (1973). "Covariance stabilizing transformations." *Annals of Statistics* 1(1).
- James, W., and Stein, C. (1961). "Estimation with quadratic loss." *Berkeley Symposium on Mathematical Statistics and Probability* 1.
- Kiefer, J., and Wolfowitz, J. (1956). "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters." *Annals of Mathematical Statistics* 27(4).
- Kline, P., Rose, E., and Walters, C. (2022). "Systemic discrimination among large US employers." *Quarterly Journal of Economics* 137(4).

References

- Kline, P., Saggio, R., and Sølvsten, M. (2020). "Leave-out estimation of variance components." *Econometrica* 88(5).
- Kling, J., Liebman, J., and Katz, L. (2007). "Experimental analysis of neighborhood effects." *Econometrica* 75(1).
- Koenker, R., and Gu, J. (2017). "REBayes: an R package for empirical Bayes mixture methods." *Journal of Statistical Software* 82(8).
- Koenker, R., and Mizera, I. (2014). "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules." *Journal of the American Statistical Association* 109(506).
- Krueger, A., and Summers, L. (1988). "Efficiency wages and the inter-industry wage structure." *Econometrica* 56(2).
- Morris, C. (1983). "Parametric empirical Bayes inference: theory and applications." *Journal of the American Statistical Association* 78(381).
- Mountjoy, J., and Hickman, B. (2021). "The returns to college(s): relative value-added and match effects in higher education." NBER working paper no. 29276.
- Narasimhan, B., and Efron, B. (2020). "deconvolveR: a G-modeling program for deconvolution and empirical Bayes estimation." *Journal of Statistical Software* 94(11).

References

- Raudenbush, S., and Bryk, A. (1986). "A hierarchical model for studying school effects." *Sociology of Education* 59(1).
- Robbins, H. (1950). "A generalization of the method of maximum likelihood: estimating a mixing distribution." *Annals of Mathematical Statistics* 21(2).
- Storey, J. (2002). "A direct approach to false discovery rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64(3).
- Storey, J. (2003). "The positive false discovery rate: a Bayesian interpretation and the *q*-value." *Annals of Statistics* 31(6).
- Storey, J., Taylor, J., and Siegmund, D. (2004). "Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66(1).