# Empirical Bayes Methods: Theory and Application

Jiaying Gu, University of Toronto
Chris Walters, UC Berkeley and NBER

NBER Methods Lectures

July 2022

# Empirical Bayes Applications

- 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)
- In settings with many unit-specific parameters, **empirical Bayes** (EB) methods are useful for
  - 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?)

# Today's Agenda

- Goals for the rest of today:
  - Recap basic EB theory
  - Illustrate through two applications
- Application 1: School value-added in Boston (Angrist, Hull, Pathak and Walters, 2017)
  - Classic parametric EB
- Application 2: Labor market discrimination among large US employers (Kline, Rose, and Walters, forthcoming)
  - Non-parametric/robust EB

# Application 1: School Value-Added

- Consider a population of students indexed by $i$, each attending one of $J$ schools in a district
- Let $Y_i(j)$ denote student i's potential academic achievement if s/he attends school $j \in \{1, ..., J\}$
- Simple additive model for potential outcomes:

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

- $\beta_j$ is the **value-added** of school $j$
- $\varepsilon_i$ represents unobserved student heterogeneity (family background, ability, etc.). Normalize $E[\varepsilon_i] = 0$
- Constant effects model: $\beta_j - \beta_k$ 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. $\beta_1$
- Might be interested in features of the *distribution* of $\beta_j$'s across schools
  - How much does school quality vary?
- Might be interested in making a decision that depends on the $\beta_j$'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:

$$Y_i = \sum_j \beta_j D_{ij} + \varepsilon_i$$

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

$$Y_i = \sum_j \beta_j D_{ij} + X_i'\gamma + u_i$$

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

# VAM Estimates

- VAM estimation yields an estimate for each school along with standard errors: $\{\hat{\beta}_j, s_j\}_{j=1}^{J}$
- Assume:

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

- Think of this as an asymptotic approximation: schools are large enough for estimates to be approximately normal and centered at the truth, with variance $\approx s_j^2$

# Introducing $G$

- Second level of the hierarchy describes the cross-school distribution of value-added:

$$\beta_j \sim G(\beta), \; j = 1, ...., J$$

- 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 value-added
  - How much does school quality vary? $\sigma_\beta^2 = \int (\beta - \mu_\beta)^2 dG(\beta)$
  - What's the difference between 75th and 25th percentiles of value-added? $G^{-1}(0.75) - G^{-1}(0.25)$
- EB **deconvolution**: Use noisy estimates $\hat{\beta}_j$ along with standard errors $s_j$ to compute an estimate $\hat{G}$ of $G$

# The Philosophy of $G$

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

# Normal/Normal Model

- Suppose $G$ is normal and independent of $s_j$
- Then we have the hierarchical model

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

$$\beta_j | s_j \sim N(\mu_\beta, \sigma_\beta^2)$$

- **Hyperparameters** $\mu_\beta$ and $\sigma_\beta^2$ summarize the value-added distribution
- With this model for $G$, deconvolution just requires estimating these two hyperparameters

# Estimating Hyperparameters

- Common estimators for value-added hyperparameters:

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

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

- Subtracting $s_j^2$ is a bias-correction accounting for excess variance in $\hat{\beta}_j$'s due to sampling error
  - $\hat{\sigma}_\beta^2 > 0 \implies$ **overdispersion** beyond what we'd expect from noise
- Other approaches: MLE; Kline, Saggio, and Sølvsten (2020) unbiased variance estimator

# Posterior Means

- In normal/normal model, posterior mean for $\beta_j$ given $(\hat{\beta}_j, s_j)$ is:

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

- Posterior mean **shrinks** noisy estimate $\hat{\beta}_j$ toward prior mean based on signal-to-noise ratio
- Linear shrinkage formula coincides with regression of $\beta_j$ on $\hat{\beta}_j \implies$ minimum mean squared error (MSE) linear predictor even if $G$ isn't normal

# EB Posterior Means

- Putting the "E" in "EB" – Empirical Bayes posterior mean $\hat{\beta}_j^*$ plugs in estimated hyperparameters $\hat{\sigma}_\beta^2$ and $\hat{\mu}_\beta$:

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

- 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)

  - "Learning from the experience of others" (Efron, 2012)

# Summary: A Three-step EB Recipe

1. **Effect estimation:** Estimate parameter for each unit $\implies \{\hat{\beta}_j, s_j\}_{j=1}^{J}$

2. **Deconvolution:** Use $\{\hat{\beta}_j, s_j\}_{j=1}^{J}$ to estimate mixing distribution $\implies \hat{G}$

3. **Posterior formation:** Treating $\hat{G}$ as prior, update with $(\hat{\beta}_j, s_j)$ to form posterior $\implies \{\hat{\beta}_j^*\}_{j=1}^{J}$

# When to Shrink?

- Should we prefer the shrunk posterior mean to the unbiased estimate $\hat{\beta}_j$? It depends on our goals
- Conditional on the value-added of school $j$, MSE for the two estimators is:

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

$$E\left[(\beta_j^* - \beta_j)^2 | \beta_j, s_j\right] = \left(\frac{\sigma_\beta^2}{\sigma_\beta^2 + s_j^2}\right)^2 s_j^2 + \left(\frac{s_j^2}{\sigma_\beta^2 + s_j^2}\right)^2 (\beta_j - \mu_\beta)^2$$

- If we're only interested in one school (e.g. $\beta_1$), not clear which is better
- Shrinkage reduces variance, but may introduce substantial bias if the school is very different from average

# When to Shrink?

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

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

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

- Linear shrinkage estimate is superior if we want an estimator that performs well on average across schools
  - Holds whether or not $G$ is normal (James/Stein 1961 result)
  - See Armstrong et al. (forthcoming) on robust inference

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

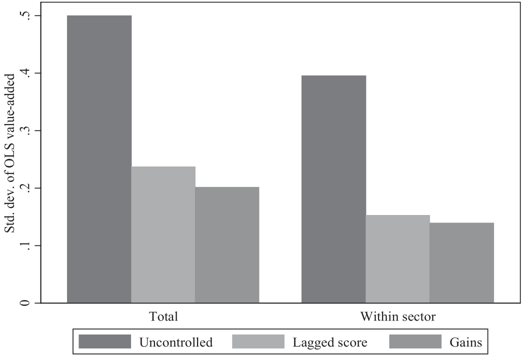


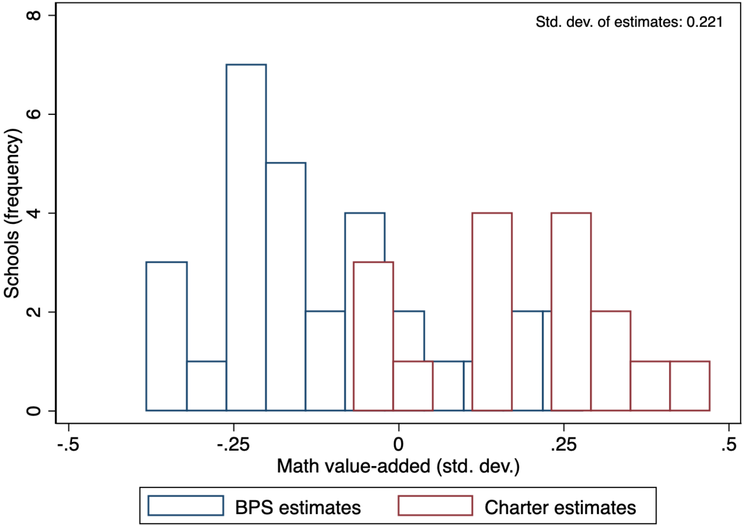

FIGURE I

Standard Deviations of School Effects from OLS Value-Added Models

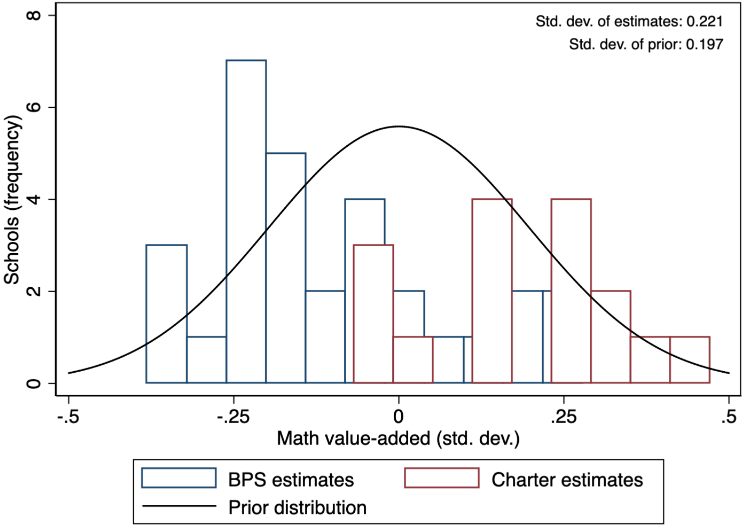
This figure compares standard deviations of school effects from alternative OLS value-added models. The notes to Table III describe the controls included in the lagged score and gains models; the uncontrolled model includes only year effects. The variance of OLS value-added is obtained by subtracting 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.

Estimates from Angrist et al. (2017)

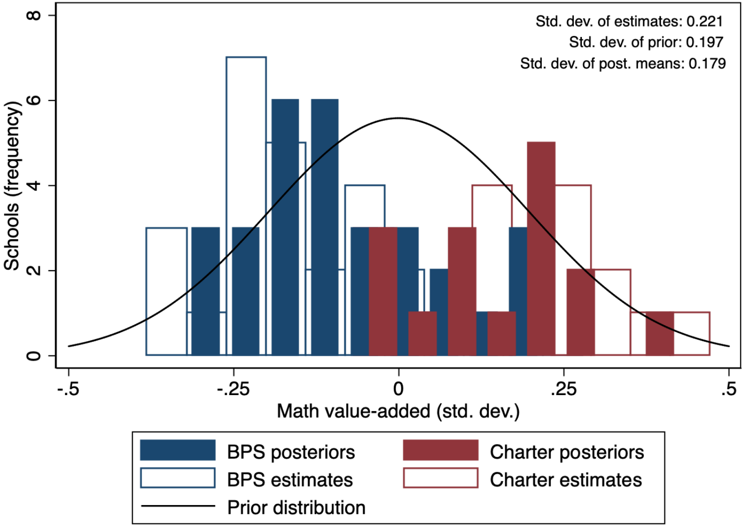
# Histogram of Lagged Score VAM Estimates for Boston (Sixth Grade Math, 2014)



# Prior Distribution Pooling Sectors



# Posterior Means Pooling Sectors



# Incorporating Covariates

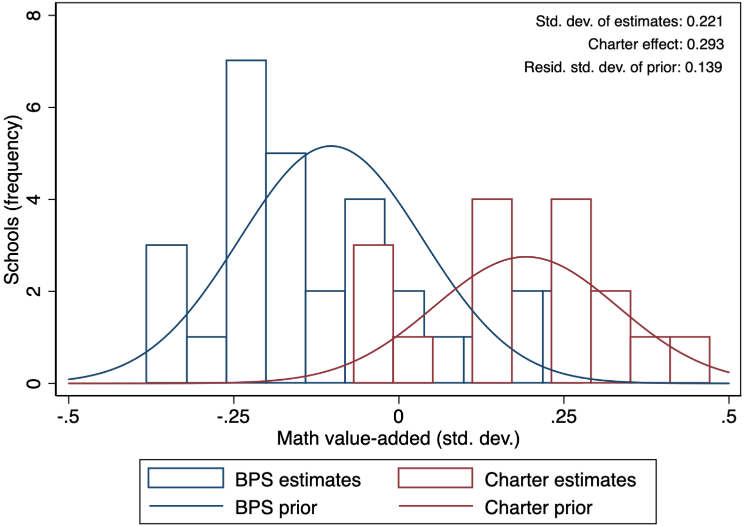
- It is often natural to build observed covariates into EB estimates
  - Learning from the experience of *which* others?
- Model for $G$ conditional on a vector of characteristics $C_j$, e.g. charter sector indicator:

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

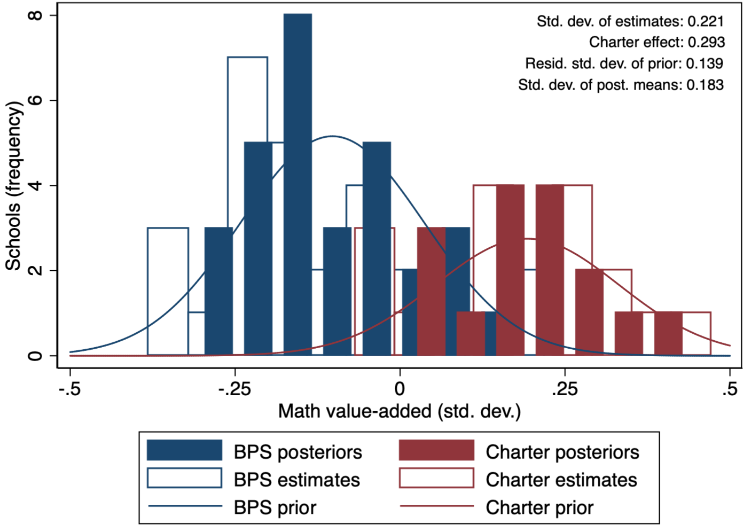
- Estimate $\mu$ from regression of $\hat{\beta}_j$ on $C_j$; deconvolve residuals $\hat{r}_j = \hat{\beta}_j - C_j'\hat{\mu}$ to estimate $\sigma_r^2$
- Resulting EB posterior shrinks $\hat{\beta}_j$ toward estimated linear index:

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

# Prior with Charter Sector Location Shift



# Posteriors Shrinking Toward Sector Means



# EB for Bias Correction

- 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 $b_j$:

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

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

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

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

# EB for Bias Correction

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

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

- We can use the ensemble $\{\hat{\alpha}_j, \hat{\beta}_j\}_{j=1}^J$ to estimate $G(\beta, b)$, the joint distribution of truth and bias
- EB "hybrid" posterior $\hat{\beta}_j^* = E_{\hat{G}}[\beta_j | \hat{\beta}_j, \hat{\alpha}_j]$ trades off bias and variance to minimize MSE:

$$\hat{\beta}_j^* = \hat{\tau}_\beta \hat{\beta}_j + \hat{\tau}_\alpha (\hat{\alpha}_j - (\hat{\mu}_\alpha - \hat{\mu}_\beta)) + (1 - \hat{\tau}_\beta - \hat{\tau}_\alpha)\hat{\mu}_\beta$$

- Angrist et al. (2017) generalize to underidentified case; see also Chetty and Hendren (2018)

# MSE Improvements from Lottery-based Hybrid Estimates

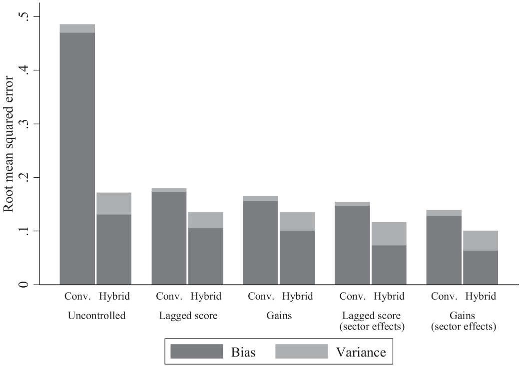



FIGURE VI

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.

# EB Decision Rules

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

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

- Cost 1 of mistakenly selecting high-performing school; cost $\kappa$ of failing to select low-performing school
- Risk-minimizing decision rule with $J$ schools:

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

# EB Decision Rules

- Solution is to select schools with sufficiently high posterior probability of value-added below $c$:

$$\delta^*(\hat{\beta}_j, s_j) = 1\left\{\Pr_G\left[\beta_j < c|\hat{\beta}_j, s_j\right] \geq \frac{1}{1+\kappa}\right\}$$

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

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

- EB decision rule plugs in estimated hyperparameters $(\hat{\mu}_\beta, \hat{\sigma}_\beta^2)$
- Different objectives call for using different functionals of posterior for decision-making
- See Gu and Koenker (2021) for EB analysis of tail selection problems

# EB and Machine Learning

- EB methods are closely related to **machine learning** (ML) approaches
- Parametric normal/normal model with $N$ students per school:

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

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

$$\beta_j \sim N(0, \sigma_\beta^2)$$

- Unbiased estimator $\bar{Y}_j = \frac{1}{N}\sum_i Y_{ij}$, with variance $Var(\bar{Y}_{ij}|\beta_j) = \sigma_\epsilon^2/N$
- Posterior distribution for $\beta_j$ is $N(\beta_j^*, V^*)$ with

$$\beta_j^* = \left(\frac{\sigma_\beta^2}{\sigma_\beta^2+\sigma_\epsilon^2/N}\right)\bar{Y}_j, \;\; V^* = \frac{\sigma_\epsilon^2\sigma_\beta^2}{N\sigma_\beta^2+\sigma_\epsilon^2}$$

# EB and Machine Learning

- Posterior density for $\beta_j$:

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

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

$$(\beta_1^*, ..., \beta_J^*) = \arg\max_{(\beta_1,...,\beta_J)} \sum_j \log f(\beta_j | Y_{1j}....Y_{Nj})$$

$$= \arg\max_{(\beta_1,...,\beta_J)} \sum_{j=1}^{J}\sum_{i=1}^{N} \log\phi\left(\frac{Y_{ij}-\beta_j}{\sigma_\epsilon}\right) + \sum_{j=1}^{J} \log\phi\left(\frac{\beta_j}{\sigma_\beta}\right) + cons$$

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

# EB and Machine Learning

- Plugging in normal density yields

$$(\beta_1^*, ..., \beta_J^*) = \arg \max_{(\beta_1,...,\beta_J)} -\sum_{j=1}^{J}\sum_{i=1}^{N}\frac{(Y_{ij}-\beta_j)^2}{2\sigma_\epsilon^2} - \sum_{j=1}^{J}\frac{\beta_j^2}{2\sigma_\beta^2}$$

$$= \arg \min_{(\beta_1,...,\beta_J)} \sum_{j=1}^{J}\sum_{i=1}^{N}(Y_{ij}-\beta_j)^2 + \frac{\sigma_\epsilon^2}{\sigma_\beta^2}\sum_{j=1}^{J}\beta_j^2$$

$$= \arg \min_{(\beta_1,...,\beta_J)} \sum_{j=1}^{J}\sum_{i=1}^{N}(Y_{ij}-\beta_j)^2 + \lambda p(\beta_1, ..., \beta_J)$$

- This is regularized least squares with an L2 (quadratic) penalty $p(\cdot)$, also known as **ridge regression**

- Empirical Bayes $\Longrightarrow$ use the data to choose tuning parameters in penalty function

# 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

# Application 2: Employer-level Labor Market Discrimination

- Kline, Rose and Walters (forthcoming) apply EB methods to study the distribution of discrimination across large US employers
- 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

# 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 $R_{ijf} \in \{b, w\}$
- Black/white 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 $\Delta_{jf}$

# Firm-level Estimates

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

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

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

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

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

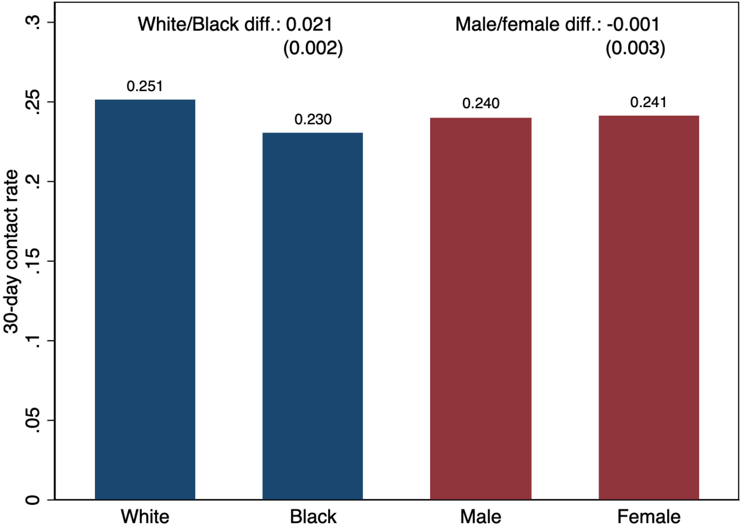
# The Distribution of Discrimination

- Let $G$ denote the distribution of contact gaps across firms:

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

- $G$ answers questions about concentration of discrimination
  - Is average white/Black difference in callbacks driven by a small share of severe discriminators?
- Start by estimating mean and variance
- Then use flexible deconvolution methods to estimate other features of $G$

# Average Contact Gaps by Race and Gender



# Variance Estimation

- Estimator for variance of $G$:

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

- Special case of unbiased leave-out variance component estimator of Kline, Saggio and Sølvsten (2020)
  - Unbiased $s_f^2 +$ degrees of freedom correction $\implies$ finite-sample unbiased estimate

- Rewrite using cross-products of job-level contact gaps:

$$\hat{\sigma}^2_\Delta = \left(\frac{F-1}{F}\right)\left[\frac{1}{F}\sum_{f=1}^{F}\frac{2}{J_f(J_f-1)}\sum_{j=2}^{J_f}\sum_{\ell=1}^{j-1}\hat{\Delta}_{fj}\hat{\Delta}_{f\ell} - \frac{2}{F(F-1)}\sum_{f=2}^{F}\sum_{k=1}^{f-1}\hat{\Delta}_f\hat{\Delta}_k\right]$$

- Interpretation: $\hat{\sigma}^2_\Delta$ measures covariance between contact gaps across jobs at the same firm

# 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 (forthcoming).

# Flexible Deconvolution

- Features of $G$ beyond the mean and variance are also of interest
- Hierarchical model:

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

$$\Delta_f \sim G(\Delta)$$

- Next, consider flexible deconvolution methods imposing little structure on $G$
- N.B.: Need to account for possible dependence between effect sizes $\Delta_f$ and sampling variance $s_f^2$
  - Maybe firms where more jobs were sampled discriminate more/less
  - Maybe firms where overall callback rates are higher discriminate more/less

# Flexible Deconvolution: Efron (2016)

- For now, sidestep precision-dependence by transforming estimates into $z$-scores
- Let $z_f = \hat{\Delta}_f / s_f$ denote the estimated $z$-score for firm $f$, and let $\mu_f = \Delta_f / s_f$ denote its population counterpart. Then

$$z_f | \mu_f \sim N(\mu_f, 1)$$

$$\mu_f \sim G_\mu(\mu)$$

- Efron (2016) proposes to approximate $G_\mu$ with distribution in smooth exponential family
  - Parameterize density with flexible spline
  - Estimate spline parameters by penalized maximum likelihood
  - Implemented in **deconvolveR** R package (Narasimhan and Efron, 2020)
  - Requires choosing penalization tuning parameter. Sensible approach: calibrate to match unbiased variance estimate

# Flexible Deconvolution: NPMLE

- Alternative approach: Non-parametric maximum likelihood estimator (NPMLE; Robbins, 1950; Kiefer and Wolfowitz, 1956)
- NPMLE picks mixing distribution to maximize likelihood of observed data:

$$\hat{G}_{\mu} = \max_{G \in \mathcal{G}} \sum_{f=1}^{F} \log \left( \int \phi \left( z_f - \mu \right) dG(\mu) \right)$$

- Solution is a discrete distribution with at most $F$ 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)
- See Koenker (2016) for a comparison of the Efron (2016) and NPMLE approaches

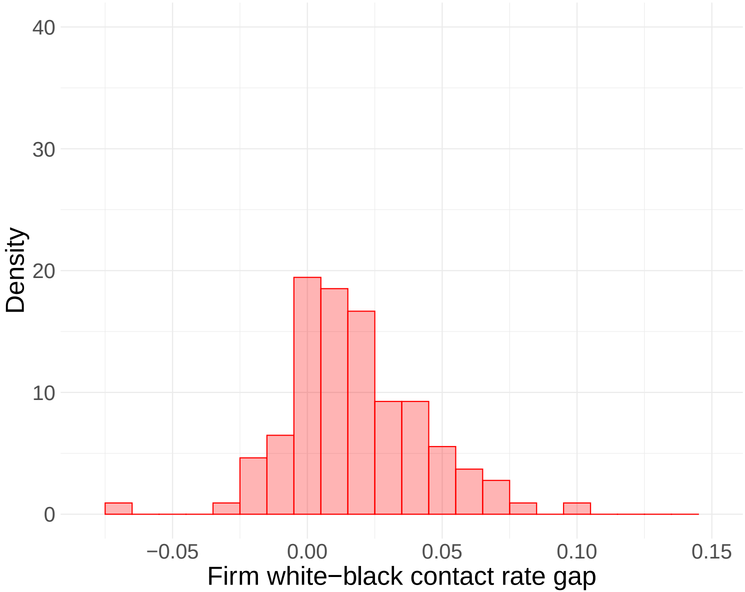
# From $z$-scores to Levels

- Suppose we have an estimate $\hat{G}_\mu$ of the distribution of $z$-scores
- To recover the distribution of $\Delta_f = \mu_f s_f$, need a change of variables
- Suppose $\mu_f$ is independent of $s_f$, and let $g_\mu$ and $h_s$ denote the densities of $\mu_f$ and $s_f$
- Density of contact gaps is then

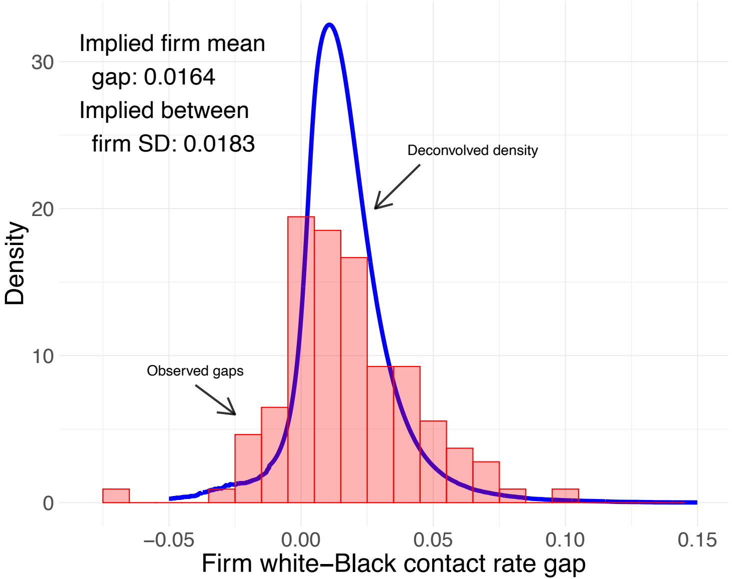
$$g_\Delta(x) = \int \frac{1}{s} g_\mu(x/s) h_s(s) ds$$

- Plug in estimated density $\hat{g}_\mu$ of $z-$scores and empirical distribution of standard errors to compute $\hat{g}_\Delta$

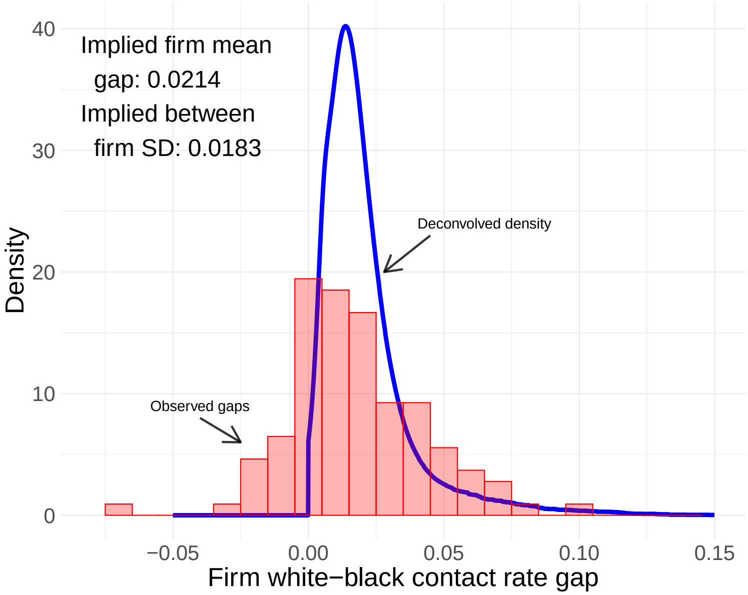
# Histogram of Race Contact Gap Estimates



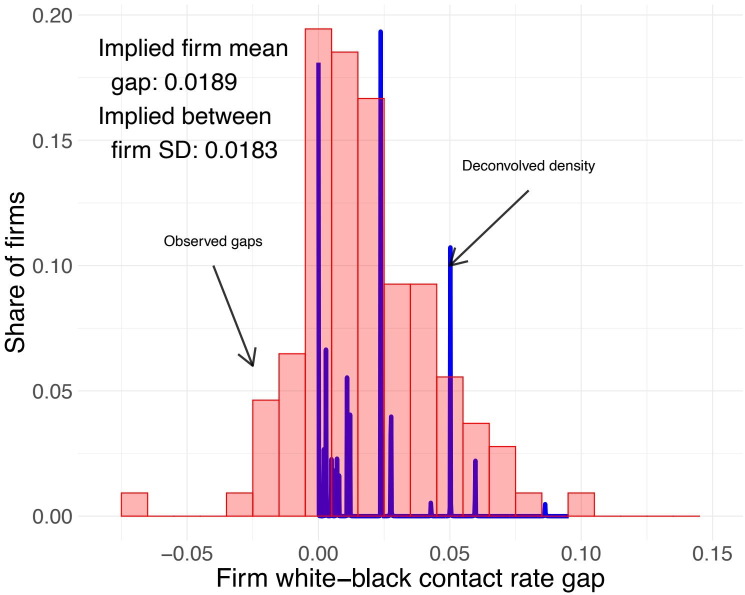
# Deconvolved Distribution of Race Contact Gaps



# Deconvolution Imposing Shape Restriction: $\Delta_f \geq 0$

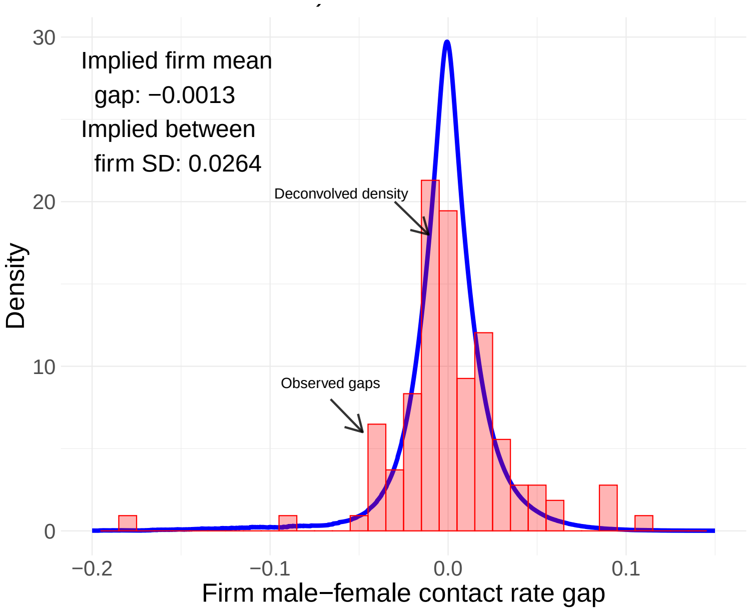


# NPMLE Deconvolution Estimates for Race

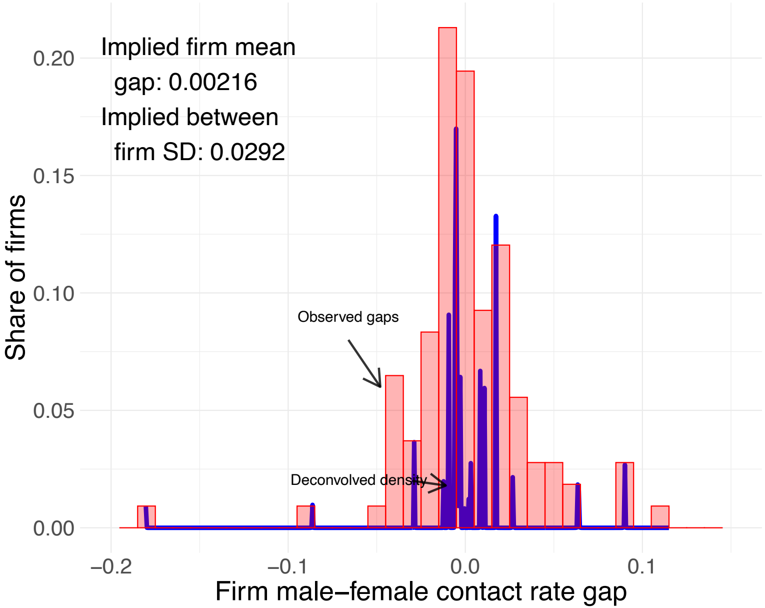


# Histogram of Gender Contact Gap Estimates

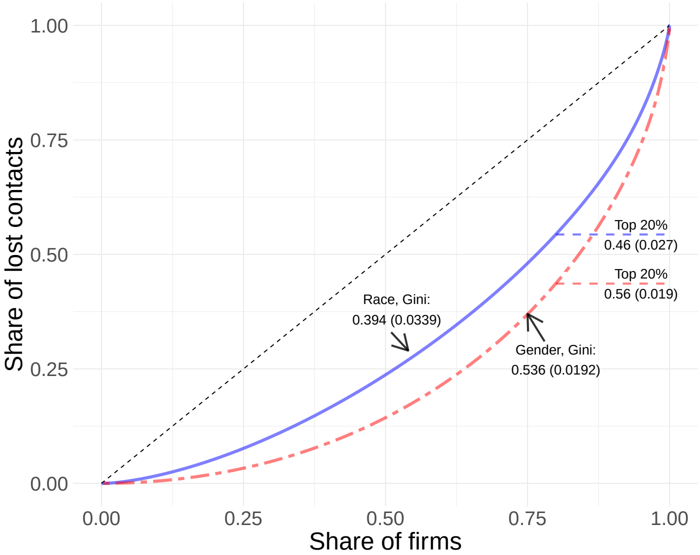
# Deconvolved Distribution of Gender Contact Gaps



# NPMLE Estimates for Gender



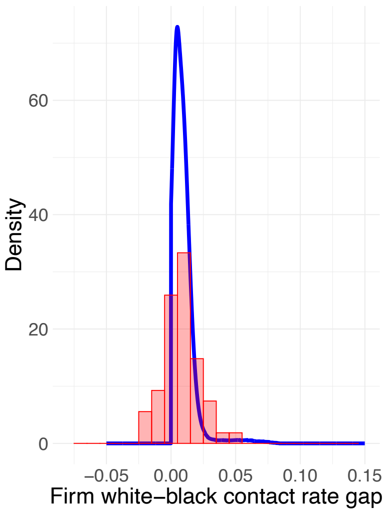
# Lorenz Curves Derived from Efron (2016) $\hat{G}$'s



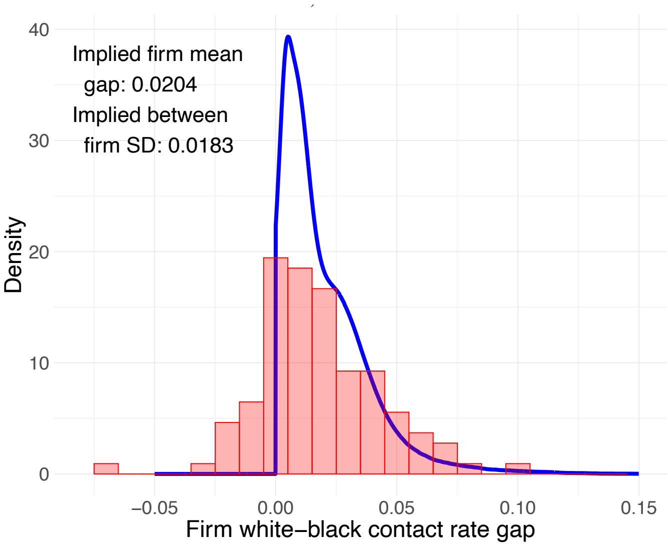
# Accounting for Precision-Dependence

- Note: if $\mu_f$ is independent of $s_f$, then effect sizes are increasing in standard errors
  - $\Delta_f = \mu_f s_f$, so $E[\Delta_f | s_f] = \bar{\mu} s_f$
  - Can test whether this approximation is reasonable
- Other approaches to dealing with dependence:
  - Treat $s_f$ as a covariate that shifts location and/or scale of $G$
  - Variance-stabilizing transformation: Find function $t(\cdot)$ such that $Var(t(\hat{\Delta}_f)|\Delta_f)$ is approximately constant (e.g. Brown, 2008)
  - Estimate bivariate distribution of $(\Delta_f, s_f)$, e.g. with NPMLE

# Separate Deconvolutions for Low vs. High $s_f$



# Marginal Distribution from Separate Deconvolutions



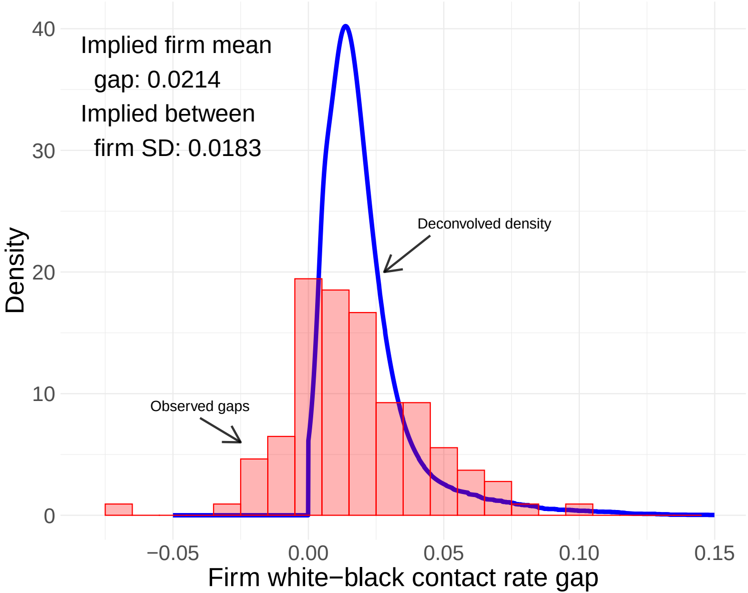
# Firm-level Posteriors

- With an estimate of the mixing distribution $\hat{G}$ in hand, move on to EB step 3: posterior estimates of firm-level discrimination
- EB posterior mean for $\Delta_f$:

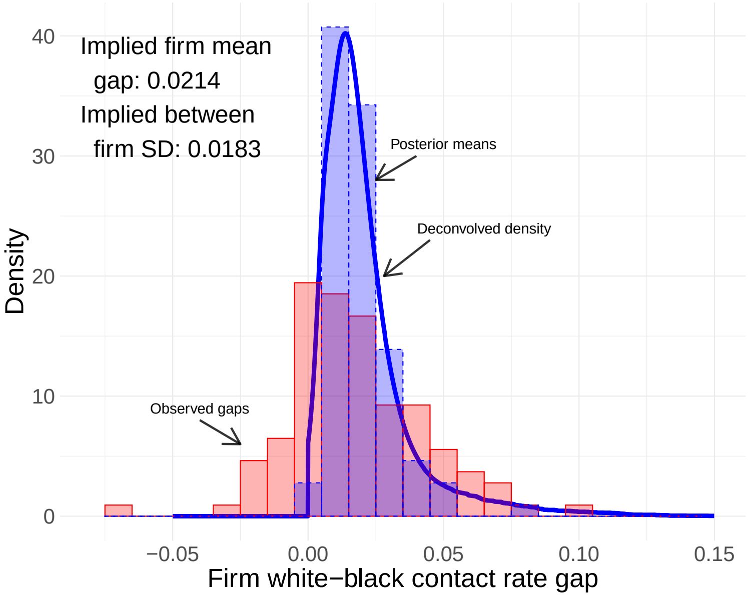
$$\hat{\Delta}_f^* = s_f \times \frac{\int x\phi(z_f - x)\hat{g}_\mu(x)dx}{\int \phi(z_f - x)\hat{g}_\mu(x)dx}$$

- Compare distributions of:
  - Unbiased estimates $\hat{\Delta}_f$
  - Contact gaps $\Delta_f$, as implied by Efron (2016) $\hat{G}$ estimate
  - EB posterior means $\hat{\Delta}_f^*$

# Distribution of Race Contact Gaps



# Histogram of Posterior Means



# Large-Scale Inference

- As with schools, we may have objectives other than minimizing MSE of discrimination estimates
- May want to make decisions about how to classify specific firms
  - Which firms are discriminating at all ($\Delta_f \neq 0$)?
  - Which firms are in the top quintile of discrimination ($\Delta_f > G^{-1}(0.8)$)?
- Such decisions are closely related to multiple-testing problems ("large-scale inference;" Efron, 2012)
- Next, consider **robust EB** methods for classifying discriminators

# Multiple Testing

- Suppose we conduct a hypothesis test for each firm, yielding a list of $p$-values $\{p_f\}_{f=1}^{F}$
- Example: one-tailed $t$-test of $H_0 : \Delta_f = 0$ vs. $H_A : \Delta_f > 0$
  - Test statistic: $z_f = \hat{\Delta}_f / s_f$
  - $P$-value: $p_f = 1 - \Phi\left(z_f\right)$
- Decision rule: reject all hypotheses with $p$-values less than $\bar{p}$
- How many mistakes do we expect to make?

# False Discovery Rates

- By Bayes rule, the expected share of non-discriminators among firms with $p$-values below $\bar{p}$ is:

$$\Pr\left[\Delta_f = 0 | p_f \leq \bar{p}\right] = \frac{\Pr\left[p_f \leq \bar{p} | \Delta_f = 0\right] \Pr[\Delta_f = 0]}{\Pr\left[p_f \leq \bar{p}\right]}$$

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

- This quantity is the **False Discovery Rate** ($FDR$) for our decision rule (Benjamini and Hochberg, 1995)

- If we can limit $FDR$ to $\bar{q}$, we should expect $100\bar{q}\%$ of firms classified as discriminators to have $\Delta_f = 0$

# Estimating *FDR*

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

- $P$-values are uniformly distributed under the null, so $\Pr\left[p_f \leq \bar{p} | \Delta_f = 0\right] = \bar{p}$
- Denominator is marginal CDF of $p$-values, estimable from empirical share below $\bar{p}$
- Difficulty is estimating $\pi_0 = \Pr[\Delta_f = 0]$, the population share of true nulls
  - $\pi_0$ is a feature of $G$: $\pi_0 = \int 1[\Delta = 0] dG(\Delta)$
  - $\pi_0$ is not point-identified: can't tell the difference between worlds where a mass of firms have $\Delta_f$ exactly 0 vs. vanishingly small
  - Efron (2016) continuous approximation automatically implies $\hat{\pi}_0 = 0$

# Bounding $\pi_0$

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

- 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 $F_p(\bar{p}) >> \bar{p}$
  - $\pi_0$ can't be 1 if mean or variance of $G \neq 0$
  - We can borrow strength from the ensemble of tests to bound $\pi_0$

# Bounding $\pi_0$

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

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

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

$$\pi_0 \leq \min_u \ f_p(u)$$

- We expect density of false nulls to be concentrated toward zero $\implies$ tightest bound near 1. Storey (2002) proposes tail-density estimator:

$$\hat{\pi}_0 = \frac{\sum_{f=1}^{F} 1\{p_f > \lambda\} p_f}{(1 - \lambda) F}$$

- Higher $\lambda$ means tighter bound but noiser estimate – Storey et al. (2004) propose bootstrap procedure to select $\lambda$
- Armstrong (2015) provides confidence interval for $\pi_0$

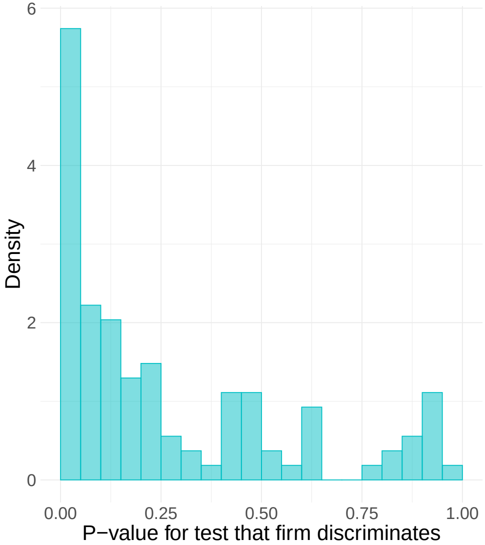
# $q$-values for FDR Control

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

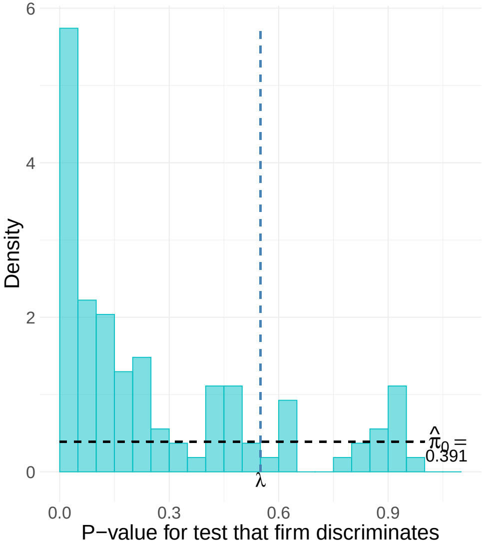
$$q_f = \widehat{FDR}(p_f) = \frac{p_f \hat{\pi}_0}{\hat{F}_p(p_f)}$$

- $q$-value $\approx$ EB equivalent of $p$-value
  - Rather than controlling $\Pr[Reject_f = 1 | \Delta_f = 0]$, use Bayes rule + ensemble of tests to control $\Pr[\Delta_f = 0 | Reject_f = 1]$
- If firm $f$'s $q$-val is $q_f$ and we reject all hypotheses with $p$-vals lower than $p_f$, we should expect *at most* $100q_f\%$ of rejections to be mistakes

# $P$-value Histogram from One-Tailed Tests of $H_0 : \Delta_f \leq 0$



# $\hat{\pi}_0 = 0.39 \implies$ At Least 61% of Firms Discriminate Against Black Applicants



# 23 of 108 Firms Have $q_f \leq 0.05$

| 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 |

# EB for Decision-Making

- What feature of posterior should we use for decisions? As usual, depends on our objectives
- Suppose an auditor is interested in investigating discriminators, with utility function

$$U(\delta) = \sum_{f=1}^{F} \delta_f \left( \Delta_f^{1/\rho} - c \right)$$

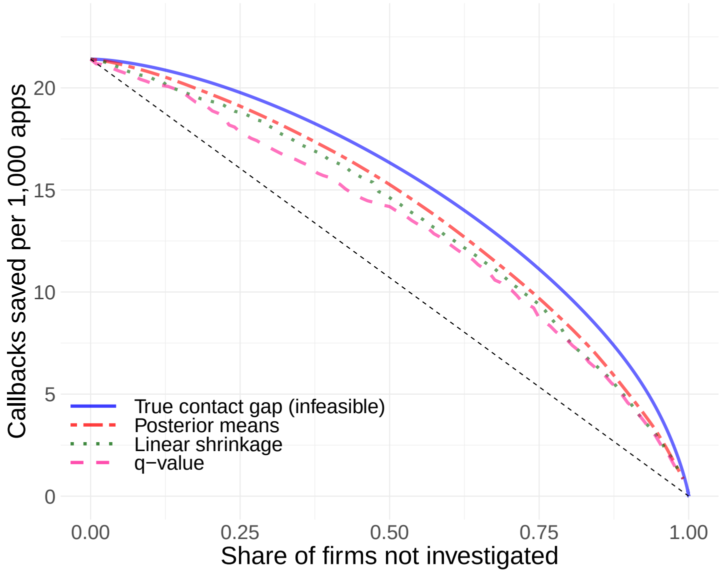
- $\delta_f \in \{0, 1\}$ is investigation indicator, $c$ is investigation cost, $\rho \geq 1$ indexes risk aversion
- With prior $G$ and evidence $\mathcal{E} = \{\hat{\Delta}_f, s_f\}_{f=1}^{F}$, expected-utility maximizing rule is:

$$\delta_f^* = 1\left\{ E_G\left[ \Delta_f^{1/\rho} | \mathcal{E} \right] > c \right\}$$

# EB for Decision-Making

- When $\rho = 1$, $\delta_f^* = 1\left\{\Delta_f^* > c\right\}$
  - Risk-neutral auditor investigates based on posterior mean
- When $\rho \to \infty$ , $\delta_f^* = 1\left\{\Pr_G\left[\Delta_f = 0|\mathcal{E}\right] < 1 - c\right\}$
  - Risk-averse auditor investigates based on **local false discovery rate** – motivates *FDR* cutoff rule
  - $q$-value decision rule motivated by optimizing against least-favorable $G$ (highest $\pi_0$) in identified set
  - See Kline and Walters (2021) for minimax approach to job-level discrimination with partial identification of $G$

# Detection Frontiers Implied by Efron (2016) $\hat{G}$



# Thanks

- Feel free to contact us with questions or issues:
  - Jiaying: jiaying.gu@utoronto.ca
  - Chris: crwalters@econ.berkeley.edu
- Data and code for employment discrimination application available online:
  - `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HLO4XC`
  - Try it out yourself!

# 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).
- Armstrong, T. (2015). "Adaptive testing on a regression function at a point." *Annals of Statistics* 43(5).
- Armstrong, T., Kolesar, M., and Plagborg-Møller (forthcoming). "Robust empirical Bayes confidence intervals." *Econometrica*.
- 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).
- Brown, L. (2008). "In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies." *Annals of Applied Statistics* 2(1).

# References

- 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).
- 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.
- 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).
- 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. (2012). <u>Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction</u>. Cambridge University Press.
- Efron, B. (2016). "Empirical Bayes deconvolution estimates." *Biometrika* 103(1).

# References


- 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).
- 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).
- Gu, J., and Koenker, R. (forthcoming). "Invidious comparisons: ranking and selection as compound decisions." *Econometrica*.
- 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. (forthcoming). "Systemic discrimination among large US employers." *Quarterly Journal of Economics*.
- Kline, P., Saggio, R., and Sølvsten, M. (2020). "Leave-out estimation of variance components." *Econometrica* 88(5).
- Kline, P., and Walters, C. (2021). "Reasonable doubt: experimental detection of job-level employment discrimination." *Econometrica* 89(2).

# References


- Kling, J., Liebman, J., and Katz, L. (2007). "Experimental analysis of neighborhood effects." *Econometrica* 75(1).
- Koenker, R. (2016). "Bayesian deconvolution: an R vinaigrette." CEMMAP working paper CWP38/17.
- 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).
- Lehmann, E., and Romano, J. (2005). "Generalizations of the familywise error rate." *Annals of Statistics* 33(3).
- 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.

# References


- Narasimhan, B., and Efron, B. (2020). "deconvolveR: a G-modeling program for deconvolution and empirical Bayes estimation." *Journal of Statistical Software* 94(11).
- 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).