ROCKWOOL FOUNDATION
BERLIN
INSTITUTE FOR THE ECONOMY AND THE FUTURE OF WORK

CReAM | Centre for Research & Analysis of Migration

DISCUSSION PAPER SERIES 22/24

Empirical Bayes Methods in Labor Economics

Christopher Walters

OCTOBER 2024

ROCKWOOL Foundation Berlin
www.rfberlin.com

Centre for Research & Analysis of Migration
www.cream-migration.org

Empirical Bayes Methods in Labor Economics

Christopher Walters*
UC Berkeley, NBER, and Amazon

October 25, 2024

Abstract

Labor economists increasingly work in empirical contexts with large numbers of unit-specific parameters. These settings include a growing number of *value-added* studies measuring causal effects of individual units like firms, managers, neighborhoods, teachers, schools, doctors, hospitals, police officers, and judges. Empirical Bayes (EB) methods provide a powerful toolkit for value-added analysis. The EB approach leverages distributional information from the full population of units to refine predictions of value-added for each individual, leading to improved estimators and decision rules. This chapter offers an overview of EB methods in labor economics, focusing on properties that make EB useful for value-added studies and practical guidance for EB implementation. Applications to school value-added in Boston and employer-level discrimination in the US labor market illustrate the EB toolkit in action.

Keywords: empirical Bayes, labor economics, value-added, shrinkage, Bayesian methods, multiple testing

*E-mail: crwalters@econ.berkeley.edu. This chapter was prepared for Volume 5 of the Handbook of Labor Economics (HOLE). The material builds on the 2022 NBER Methods Lecture "Empirical Bayes Methods: Theory and Application." I thank Jiaying Gu for her collaboration on that lecture as well as course participants for engaging comments and questions. I also thank participants in several related workshops including the AEA Continuing Education Program, Bonn/Mannheim CRC Summer School, Northwestern Causal Inference Workshop, Empirical Bayes Mixtape Session, and UCLA CCPR seminar. I am grateful to Thomas Lemieux and Christian Dustmann for their work organizing the HOLE Volume, to participants at the 2023 HOLE Volume Conference, and to the Rockwool Foundation Berlin for providing funding for the conference. The Massachusetts Department of Elementary and Secondary Education generously provided the data used for school value-added analysis. Finally, I thank Joshua Angrist, Peter Hull, Patrick Kline, Parag Pathak, and Evan Rose for collaborations and discussions that were essential to the development of this chapter.

1 Introduction

Labor economists in a variety of research areas increasingly drill down to study large numbers of finely-grained, unit-specific parameters. Large-scale administrative data sets and new research designs have allowed researchers to sharpen the focus of scientific inquiry into economic impacts of specific institutions and individuals. In the economics of education, for example, a classic topic is the return to education, traditionally defined as the causal effect of an extra year of schooling on earnings (Card, 1999). This question has evolved into studies of the effects of attending more versus less selective colleges (Dale and Krueger, 2002, 2014) and impacts of specific colleges on labor market success (Mountjoy and Hickman, 2021; Chetty et al., 2023). Building on Krueger and Summers' (1988) seminal study of industry-specific wage premia, a large recent literature looks at the role of individual firms in wage determination (Abowd et al., 1999; Card et al., 2013, 2018; Bonhomme et al., 2023). More broadly, many areas of applied microeconomics have seen the rise of *value-added* studies considering causal effects of individual units such as neighborhoods, teachers, schools, managers, doctors, health insurance plans, hospitals, nursing homes, police officers, and judges (Chetty and Hendren, 2018; Chetty et al., 2018, 2014a; Angrist et al., 2017, 2024a; Fenizia, 2022; Chan et al., 2022; Abaluck et al., 2021; Einav et al., 2022; Goncalves and Mello, 2021; Frandsen et al., 2023).

Empirical Bayes (EB) methods provide a powerful suite of econometric tools for settings with large numbers of unit-specific parameters. The EB framework, pioneered by Robbins (1951; 1956; 1964), is designed for contexts involving multiple parallel estimation problems for many similar units. EB leverages this common structure by pooling information on all units to estimate a distribution of parameters in the population being studied. This estimated distribution is then used as an empirical prior to construct posterior predictions for each individual unit. On average, the resulting EB estimators and decision rules often perform better than approaches that consider each unit separately.

An EB approach can be used to accomplish several key objectives in value-added analyses. First, the estimated EB prior characterizes the *distribution* of value-added across units. This condenses large sets of value-added estimates into interpretable summaries of heterogeneity, such as the variance of value-added. Second, by "borrowing strength" from other similar units via the estimated prior (Morris, 1983), EB improves estimates of individual value-added parameters. Third, EB methods are useful for making decisions. EB is intimately linked to a *compound decision problem* in which an analyst faces repeated decisions across many units and seeks to minimize an aggregate loss function. A clear view of the loss function allows a researcher to convey the information about each unit's value-added that is most relevant for the economic objective at hand.

This chapter offers an overview of empirical Bayes methods in labor economics. Many excellent surveys of EB and related methods are already available (see, e.g., Efron, 2012; Bonhomme and Denis, 2024; Koenker and Gu, 2024). I do not aim to improve upon the technical elements of these treatments or break new theoretical ground on the properties of EB. My goal is to provide an

accessible toolkit for labor economists using EB methods in value-added studies. Along the way I will touch on practical issues that arise in EB analyses of microeconomic data, offer concrete examples and guidance on EB implementation, and make connections between EB and other methods that may be more familiar to labor economists. Applications to school value-added in Boston and employer-level discrimination in the US labor market will illustrate the EB toolkit in action.

The remainder of the chapter is organized in two broad sections. Section 2 provides a basic introduction to the empirical Bayes approach. I start with a simple three-step EB recipe: (1) estimate a parameter and associated standard error for each unit; (2) pool the estimates and standard errors to estimate the prior distribution (a procedure known as *deconvolution*); and (3) use the estimated prior to generate posterior predictions for each unit (also known as *empirical Bayes shrinkage*). Applying this recipe with a normal prior distribution gives rise to *linear shrinkage* estimators of the sort considered by James and Stein (1961). While linear shrinkage is nominally based on a normality assumption, the James/Stein Theorem establishes that this method improves aggregate mean squared error (MSE) even if the normal model is wrong. Thus, we can view parametric empirical Bayes as a device for motivating procedures that perform well for a broader class of data generating processes.

After covering the basics of linear shrinkage I consider several variations on the theme. These extensions include the use of shrinkage in a regression context, posteriors that incorporate observed unit characteristics, and scenarios with multiple parameters per unit or multiple research designs for estimating the same parameter. I also discuss EB decision rules for objectives other than mean squared error, outline approaches to incorporating dependence between effect sizes and standard errors, and make connections between EB shrinkage and commonly-used machine learning algorithms. Section 2 ends with an EB analysis of middle school value-added in Boston based on data from Angrist et al. (2017).

Section 3 introduces non-parametric empirical Bayes methods that relax the normality and independence assumptions maintained earlier in the chapter. Here I link the EB framework with recent approaches to bias-corrected variance component estimation (e.g., Kline et al., 2020), which may be used to estimate the variance of value-added under weak assumptions. I then consider non-parametric methods for estimating the full prior distribution, including non-parametric maximum likelihood (NPMLE; Robbins 1950; Kiefer and Wolfowitz 1956) and flexible smooth deconvolution estimators (Efron, 2016). These non-parametric prior estimation procedures give rise to corresponding non-parametric EB posteriors, which generalize James/Stein-style shrinkage to account for features of the value-added distribution beyond the mean and variance.

The second half of Section 3 covers variants of non-parametric EB that are relevant for empirical practice. These include cases with partial identification of priors and posteriors, EB approaches to multiple testing, and ranking problems. I connect ideas from throughout the chapter in a non-parametric version of the compound decision framework, and contrast EB decision rules motivated by different economic objectives. Revisiting a labor market correspondence experiment studied by Kline, Rose and Walters (2022; 2024), I apply non-parametric EB methods to flexibly estimate

distributions of race and gender discrimination across large US employers.[1] The chapter concludes in Section 4 with thoughts on directions for future applications of empirical Bayes methods in labor economics.

2 Empirical Bayes Basics

2.1 An Empirical Bayes Recipe

Consider a hierarchical data structure with individuals nested within groups. We observe data Y_i on N individuals indexed by i, each associated with one of J groups (I refer to the groups interchangeably as *units*). Let $D_i \in \{1, ..., J\}$ denote the group for individual i. A group-specific parameter θ_j determines the distribution of Y_i for individuals with $D_i = j$. For example, Y_i might represent test scores for students nested within schools, with θ_j the causal effect of school j on student achievement; or Y_i might measure log earnings for workers nested within firms, with θ_j a pay premium associated with working at firm j. I next outline a three-step empirical Bayes recipe for estimating the unit-specific θ_j parameters, characterizing the distribution of these parameters across units, and using this distributional information to refine the estimate for each unit.

Step 1: Estimation

The first step of an EB analysis uses the data for group j to form an estimate $\hat{\theta}_j$ of the parameter θ_j along with a corresponding standard error s_j. In much of what follows I will assume these estimates are unbiased, normally distributed, and mutually independent, with sampling variances equal to their squared standard errors:

$$\hat{\theta}_j | \theta_j, s_j \sim \mathcal{N}(\theta_j, s_j^2). \tag{1}$$

The conditioning on (θ_j, s_j) in this expression is to emphasize that at this stage these quantities are treated as fixed parameters rather than random variables. Normality of $\hat{\theta}_j$ may be justified either by a parametric model for the microdata Y_i or by an asymptotic approximation with a growing number of individuals in each group. In the latter case a central limit theorem is typically invoked to argue that $\hat{\theta}_j$ is asymptotically normal with limiting variance s_j^2. I later discuss scenarios where such asymptotic approximations break down.

Two examples of the estimation step will help fix ideas as I develop the EB recipe.

Example 1: Normal means

Suppose the data Y_i are normally distributed with a group-specific mean and variance:

$$Y_i | D_i = j, \theta_j, \sigma_j \sim \mathcal{N}(\theta_j, \sigma_j^2). \tag{2}$$

The natural estimator of θ_j in this case is the sample mean $\hat{\theta}_j = n_j^{-1} \sum_{i=1}^N D_{ij} Y_i$, where $D_{ij} = 1\{D_i = j\}$ indicates that observation i is in group j and $n_j = \sum_{i=1}^N D_{ij}$ is the number of observations in this group. The parametric model in (2) implies $\hat{\theta}_j$ is normally distributed with mean θ_j and

[1] Software for implementing these methods is available in this chapter's replication package.

variance $s_j^2 = \sigma_j^2/n_j$.

Example 2: School value-added

Consider a causal model of school effectiveness in which potential academic achievement for student i if she attends school j is given by

$$Y_i(j) = \theta_j + a_i. \tag{3}$$

Here θ_j represents the causal contribution of school j to student achievement (school j's *value-added*), while a_i captures all student-specific factors that influence achievement including family background, earlier educational investments, and innate ability. I normalize $E[a_i] = 0$ so that $\theta_j = E[Y_i(j)]$ equals the population mean potential outcome for school j. The additive structure in equation (3) implies school value-added is constant across students – for any student i, $\theta_j - \theta_k$ is the causal effect of attending school j rather than k. The observed outcome Y_i is the potential outcome associated with the school that i attends: $Y_i = \sum_{j=1}^{J} D_{ij} Y_i(j)$.

School enrollment is not randomly assigned, so uncontrolled comparisons of test scores across schools are likely to be contaminated by differences in student ability. This motivates an empirical strategy that adjusts for observed characteristics such as demographics and lagged achievement. Let X_i represent a vector of these control variables, de-meaned so that $E[X_i] = 0$. Write the linear projection of a_i on X_i as:

$$a_i = X_i'\gamma + \epsilon_i, \tag{4}$$

where $\gamma = E[X_i X_i']^{-1} E[X_i a_i]$, and $E[\epsilon_i] = E[X_i \epsilon_i] = 0$ by definition of γ. I can then write the observed outcome for student i as

$$Y_i = \sum_{j=1}^{J} \theta_j D_{ij} + X_i'\gamma + \epsilon_i. \tag{5}$$

The residual ϵ_i is uncorrelated with X_i by construction but need not be uncorrelated with the school indicators D_{ij}. Studies of such *value-added models* (VAMs) typically proceed under the assumption that $E[D_{ij}\epsilon_i] = 0 \ \forall j$. This is a selection-on-observables restriction requiring additive controls for X_i to eliminate any relationship between potential outcomes and school attendance. Under this assumption, Ordinary Least Squares (OLS) estimation of (5) using a random sample of students recovers unbiased estimates $\hat{\theta}_j$ of the causal value-added parameters θ_j. The squared standard errors s_j^2 are the first J diagonal elements of the asymptotic covariance matrix of the OLS coefficient vector $(\hat{\theta}_1, ..., \hat{\theta}_J, \hat{\gamma}')'$, which may be estimated using the standard White (1980) heteroskedasticity-robust variance matrix or another appropriate variance estimator. With a sufficient number of students per school we can treat $\hat{\theta}_j$ as normally distributed with variance approximately equal to s_j^2.

Step 2: Deconvolution

The second step of the EB recipe is predicated on a model of the group-level parameters as random draws from a probability distribution. Suppose the θ_j's are generated by independent draws from a common cumulative distribution function G:

$$\theta_j \sim G,\ j \in \{1, ..., J\}. \tag{6}$$

The *mixing distribution* G is central to the empirical Bayes approach. Equations (1) and (6) imply the estimates $\hat{\theta}_j$ are a mixture of draws from G and normally-distributed sampling error. In other words, the marginal distribution of $\hat{\theta}_j$ is a *convolution* of G and normal noise with variance s_j^2.

Though G will play the role of a Bayesian prior in much of what follows, it is worth highlighting that this distribution represents an objective description of the variation in parameters θ_j rather than subjective beliefs about their likely values. In the school value-added context (Example 2), we might wonder whether there are substantial differences between schools' contributions to student learning. The variance of the mixing distribution, given by $\sigma_\theta^2 = \int (\theta - \mu_\theta)^2 dG(\theta)$ with $\mu_\theta = \int \theta dG(\theta)$, provides a quantitative answer to this question. Likewise, the gap in effectiveness between the most- and least-effective schools is a feature of G; the difference in value-added between 90th and 10th-percentile schools is $G^{-1}(0.9) - G^{-1}(0.1)$. The second step of an EB analysis answers such questions empirically via *deconvolution:* extract an estimate $\hat{G}$ of the mixing distribution G from the noisy group-specific estimates $\hat{\theta}_j$ and standard errors s_j, and use $\hat{G}$ to quantify heterogeneity in parameters across groups.

While there is nothing subjective about G, a practical empiricist might nonetheless be uncomfortable with the random effects model in (6). On a conventional fixed effects view, the parameters θ_j are just a list of J unknown numbers, so it may seem unnatural to treat them as random draws from a probability distribution. In some cases model (6) is justified by a hierarchical sampling process – perhaps the groups were drawn at random from a larger superpopulation, as in some multi-site randomized trials (see, e.g., Walters, 2015). In other cases, however, we observe all the relevant groups, so it is unattractive to appeal to such a superpopulation view. In a school value-added analysis for a particular district like Boston, where are the missing schools that might have otherwise been sampled?

An important theme of this chapter is that EB methods perform well even judged on purely frequentist terms, independent of such philosophical considerations. As discussed further below, estimators and decision rules incorporating distributional information based on an estimate of G can improve on approaches treating each $\hat{\theta}_j$ in isolation. These improvements are achieved even if the stylized model in (6) is wrong. We can therefore think of G as a device for motivating procedures with desirable frequentist properties, irrespective of whether the parameters are fixed or random. In what follows I will often make use of continuous models for G, which must be misspecified on a fixed effects view; for a committed frequentist, these models may be seen as useful approximations rather than literal descriptions of the data-generating process.

To illustrate the basic mechanics of EB, the remainder of this section will mostly focus on a parametric normal model for G:

$$\theta_j | s_j \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2). \tag{7}$$

Model (7) is written conditional on s_j, which implies the effect sizes θ_j are independent of the sampling variances s_j^2 across groups, a restriction I relax later. With this parametric model deconvolution amounts to estimating the two *hyperparameters* μ_θ and σ_θ. The "hyperparameter" label refers to the fact that μ_θ and σ_θ govern the distribution of lower-level parameters (the θ_j's). Simple estimators for these hyperparameters are given by:

$$\hat{\mu}_\theta = \frac{1}{J}\sum_{j=1}^{J} \hat{\theta}_j, \tag{8}$$

$$\hat{\sigma}_\theta^2 = \frac{1}{J}\sum_{j=1}^{J} \Big[(\hat{\theta}_j - \hat{\mu}_\theta)^2 - s_j^2\Big]. \tag{9}$$

The subtraction of s_j^2 in (9) is a *bias-correction* that accounts for excess variability of the $\hat{\theta}_j$'s due to statistical noise. We should expect some chance variation in the estimated $\hat{\theta}_j$'s even if all the latent θ_j's are equal, so the naive sample variance of $\hat{\theta}_j$'s is too large relative to the variance of the mixing distribution. Subtracting the average squared standard error removes this excess noise. A finding that $\hat{\sigma}_\theta^2 > 0$ indicates *overdispersion* in the $\hat{\theta}_j$'s: the observed estimates differ more than should be expected from sampling error, implying variation in the underlying θ_j's. With a growing number of groups J, the bias-corrected variance estimate $\hat{\sigma}_\theta^2$ is consistent for the hyperparameter σ_θ^2.[2] Section 3 considers unbiased estimation of the mixing variance in finite samples and nonparametric deconvolution methods for flexibly estimating the hyperparameters of a non-normal G.

Step 3: Shrinkage

The third and final step of an EB analysis uses the noisy $\hat{\theta}_j$'s together with the estimated mixing distribution $\hat{G}$ to form posteriors for each θ_j. Specifically, we treat the deconvolved $\hat{G}$ as a prior, then use Bayes' rule to perform an update based on $(\hat{\theta}_j, s_j)$ and generate a posterior distribution for θ_j. We are often interested in a particular feature of this posterior distribution such as the posterior mean. The use of an estimated $\hat{G}$ as a prior when forming posterior predictions is commonly known as *empirical Bayes shrinkage*.

Suppose first that the mixing distribution G is known. Equations (1) and (6) along with Bayes' rule imply the cumulative distribution function for θ_j conditional on $(\hat{\theta}_j, s_j)$ evaluated at a point t

[2] Standard errors for the hyperparameter estimates in the normal/normal model can be calculated as $SE(\hat{\mu}_\theta) = J^{-1}\sqrt{\sum_j (\hat{\theta}_j - \hat{\mu}_\theta)^2}$ and $SE(\hat{\sigma}_\theta^2) = J^{-1}\sqrt{2\sum_j (\hat{\sigma}_\theta^2 + s_j^2)^2}$.

is given by:

$$\Pr\left[\theta_j \leq t|\hat{\theta}_j, s_j\right] = \frac{\int_{-\infty}^{t} \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j-\theta}{s_j}\right) dG(\theta)}{\int_{-\infty}^{\infty} \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j-\theta}{s_j}\right) dG(\theta)} \equiv \mathcal{P}(t|\hat{\theta}_j, s_j; G). \tag{10}$$

I will sometimes refer to $\mathcal{P}(t|\hat{\theta}, s_j; G)$ as an *oracle* posterior distribution because it coincides with the posterior beliefs of an oracle who knows the mixing distribution G *a priori*. With the normal mixing distribution in model (7), this oracle posterior distribution is also normal:

$$\theta_j|\hat{\theta}_j, s_j \sim \mathcal{N}(\theta_j^*, V_j^*), \tag{11}$$

$$\theta_j^* = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\sigma_\theta^2 + s_j^2}\right)\mu_\theta, \tag{12}$$

$$V_j^* = \frac{\sigma_\theta^2 s_j^2}{\sigma_\theta^2 + s_j^2}. \tag{13}$$

The posterior mean θ_j^* *shrinks* the noisy estimate $\hat{\theta}_j$ towards the prior mean μ_θ in proportion to its signal-to-noise ratio. When $\hat{\theta}_j$ is completely uninformative or the prior distribution is degenerate ($s_j^2 \to \infty$ or $\sigma_\theta^2 \to 0$) the posterior mean equals the prior mean μ_θ. As the precision of $\hat{\theta}_j$ or the variability of the prior grow large ($s_j^2 \to 0$ or $\sigma_\theta^2 \to \infty$) the posterior mean approaches the estimate $\hat{\theta}_j$. In between these extremes, θ_j^* is a convex weighted average of the unbiased estimate $\hat{\theta}_j$ and the prior mean μ_θ.

An analyst that does not know G cannot calculate the oracle posterior formulas in equations (10)-(13). The empirical Bayes approach is to plug in an estimate $\hat{G}$ of the mixing distribution, resulting in an empirically feasible posterior:

$$\mathcal{P}(t|\hat{\theta}_j, s_j; \hat{G}) = \frac{\int_{-\infty}^{t} \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j-\theta}{s_j}\right) d\hat{G}(\theta)}{\int_{-\infty}^{\infty} \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j-\theta}{s_j}\right) d\hat{G}(\theta)}, \tag{14}$$

where $\hat{G}$ is the deconvolution estimate from step 2. With the normal model in (7), this means substituting the estimated hyperparameters $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$ from equations (8)-(9) into the posterior formulas in equations (12)-(13). This substitution yields an empirical Bayes posterior mean:

$$\hat{\theta}_j^* = \left(\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\hat{\sigma}_\theta^2 + s_j^2}\right)\hat{\mu}_\theta. \tag{15}$$

The EB posterior mean $\hat{\theta}_j^*$ shrinks the unbiased estimate $\hat{\theta}_j$ for group j using distributional information based on the ensemble of estimates for all J groups, which enters through the estimated hyperparameters $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$. Morris (1983) refers to this adjustment as "borrowing strength from the ensemble." EB shrinkage refines the estimate for each individual unit by interpreting it in the context of results for a larger pool units that are similar in some relevant sense. Efron (2012) labels

such refinements "learning from the experience of others."

In the remainder of this chapter I will often refer to the estimator in equation (15) as a *linear shrinkage estimator* to distinguish it from EB posterior means derived from more elaborate non-normal models for G. While linear shrinkage is motivated by normality, the resulting predictions are likely to have good properties even when G is not normal. In particular, θ_j^* coincides with the fitted value from an infeasible ordinary least squares regression of the unobserved θ_j on $\hat{\theta}_j$.[3] By standard properties of OLS regression, we can then think of the linear shrinkage estimate as a best linear approximation to the true (possibly nonlinear) conditional mean of θ_j given $\hat{\theta}_j$. As I discuss next, linear approximations of this form turn out to improve upon the unbiased estimates $\hat{\theta}_j$ (in a particular sense) regardless of the form of G.

Recap: A three-step EB recipe

The standard progression of an EB analysis is summarized in the following three-step recipe:

1. **Estimation:** Compute an estimate $\hat{\theta}_j$ and corresponding standard error s_j for each unit j.
2. **Deconvolution:** Use the estimates and standard errors $\{\hat{\theta}_j, s_j\}_{j=1}^J$ to compute an estimate $\hat{G}$ of the mixing distribution.
3. **Shrinkage:** Treating $\hat{G}$ as a prior, update with $(\hat{\theta}_j, s_j)$ to form posterior predictions $\hat{\theta}_j^*$ for each unit.

The key outputs of the analysis are typically a summary of parameter heterogeneity based on $\hat{G}$ from the deconvolution step along with posterior predictions for each unit generated in the shrinkage step.

2.2 Gains From Shrinkage

MSE improvements in the normal/normal model

In what sense do the EB posterior means $\hat{\theta}_j^*$ improve upon the unbiased estimates $\hat{\theta}_j$? I first explore this question in the context of the normal/normal model defined by equations (1) and (7), focusing on a mean squared error (MSE) criterion.[4] Assume the hyperparameters μ_θ and σ_θ^2 are known, and consider the MSE of the oracle posterior mean θ_j^* and the unbiased estimate $\hat{\theta}_j$ conditional on the unknown parameter θ_j and standard error s_j:

$$E\left[(\hat{\theta}_j - \theta_j)^2 | \theta_j, s_j\right] = s_j^2, \tag{16}$$

$$E\left[(\theta_j^* - \theta_j)^2 | \theta_j, s_j\right] = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right)^2 s_j^2 + \left(\frac{s_j^2}{\sigma_\theta^2 + s_j^2}\right)^2 (\theta_j - \mu_\theta)^2 . \tag{17}$$

[3] When θ_j is independent of s_j we have $\theta_j^* = \left(\frac{Cov(\theta_j, \hat{\theta}_j | s_j)}{Var(\hat{\theta}_j | s_j)}\right)\hat{\theta}_j + \left(1 - \frac{Cov(\theta_j, \hat{\theta}_j | s_j)}{Var(\hat{\theta}_j | s_j)}\right)\mu_\theta$.

[4] See Angrist et al. (2023) for related discussion in the context of school value-added models.

Since $\hat{\theta}_j$ is unbiased, its MSE is given by its sampling variance s_j^2. The posterior mean θ_j^* shrinks $\hat{\theta}_j$ toward a constant μ_θ, thereby reducing variance in exchange for an increase in bias. This results in a squared bias term in the conditional MSE formula, reflected in the second term in (17).

These expressions show that if an analyst is only interested in one unit (say θ_1), it is not clear which of the two estimators is better. The linear shrinkage posterior is less variable than the unbiased estimate, but for any particular unit this variance reduction may be outweighed by increased bias. Moreover, equation (17) shows that the conditional MSE of θ_j^* is not uniform in the true parameter θ_j. The bias introduced by shrinkage is worse for units that are more atypical in the sense of having θ_j's farther from the mean μ_θ. This bias can be arbitrarily large if the mixing distribution has unbounded support. An analyst interested in a specific unit and concerned about worst-case MSE might reasonably prefer the unbiased estimate $\hat{\theta}_j$.

Next, consider an analyst who cares about reporting estimates with low MSE for many units simultaneously. This analyst expects to study many units drawn from G and wants to do well on average across all of them. We can evaluate the performance of $\hat{\theta}_j$ and θ_j^* for this purpose by integrating conditional MSE over the mixing distribution:

$$E\left[(\hat{\theta}_j - \theta_j)^2 | s_j\right] = \int E\left[(\hat{\theta}_j - \theta_j)^2 | \theta_j = \theta, s_j\right] dG(\theta) = s_j^2, \tag{18}$$

$$E\left[(\theta_j^* - \theta_j)^2 | s_j\right] = \int E\left[(\theta_j^* - \theta_j)^2 | \theta_j = \theta, s_j\right] dG(\theta) = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right) s_j^2. \tag{19}$$

This integration averages the squared bias of θ_j^* over values of θ_j while leaving MSE for $\hat{\theta}_j$ unchanged.

The signal-to-noise ratio $\sigma_\theta^2/(\sigma_\theta^2 + s_j^2)$ is less than one, so equations (18) and (19) establish that unconditional MSE for the shrinkage estimator is below that of the unbiased estimator. While shrinkage increases conditional MSE for atypical values of θ_j far from the mean, it improves MSE for values close to the mean, and on average over repeated draws from G the θ_j's cannot all be atypical. The reduction in variance therefore outweighs the increased conditional bias on average, reducing overall MSE. In fact, since θ_j^* is the conditional mean of θ_j given $(\hat{\theta}_j, s_j)$, it must have lowest average MSE of all functions of $(\hat{\theta}_j, s_j)$ under the normal/normal model. As long as the hyperparameter estimates $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$ are sufficiently precise we should expect the EB posterior mean $\hat{\theta}_j^*$ to inherit the properties of the oracle posterior mean θ_j^* and improve average MSE relative to $\hat{\theta}_j$.

The James/Stein Theorem

I next show that the gains from EB shrinkage apply more generally outside the normal random effects setup. This is a classic result due to James and Stein (Stein, 1956; James and Stein, 1961), referred to here as the *James/Stein Theorem*. Suppose each $\hat{\theta}_j$ is unbiased for the corresponding θ_j and normally distributed with common sampling variance of $s_j^2 = s^2 \ \forall j$:

$$\hat{\theta}_j|\theta_j \sim \mathcal{N}(\theta_j, s^2). \tag{20}$$

We are interested in finding an estimator with low total MSE summed over all J units, treating the θ_j's as fixed but unknown parameters. For any estimator δ_j, this objective is given by:

$$MSE_\delta = \sum_{j=1}^{J} E\left[(\delta_j - \theta_j)^2|\theta_j\right]. \tag{21}$$

By equation (16), total MSE of the unbiased estimates $\hat{\theta}_j$ is equal to Js^2:

$$MSE_{\hat{\theta}} = \sum_{j=1}^{J} E\left[(\hat{\theta}_j - \theta_j)^2|\theta_j\right] = Js^2. \tag{22}$$

Now consider an alternative estimator that shrinks each $\hat{\theta}_j$ toward a constant μ as follows:

$$\hat{\theta}_j^{JS} = \left(1 - \frac{(J-2)s^2}{\sum_{k=1}^{J}(\hat{\theta}_k - \mu)^2}\right)\hat{\theta}_j + \left(\frac{(J-2)s^2}{\sum_{k=1}^{J}(\hat{\theta}_k - \mu)^2}\right)\mu. \tag{23}$$

This estimator turns out to improve MSE relative to the unbiased $\hat{\theta}_j$'s whenever the number of units J is at least 3. Assuming $J \geq 3$, we have

$$\begin{aligned} MSE_{\hat{\theta}^{JS}} &= \sum_{j=1}^{J} E\left[(\hat{\theta}_j^{JS} - \theta_j)^2|\theta_j\right] \\ &\leq Js^2 - \frac{(J-2)^2 s^4}{(J-2)s^2 + \sum_{j=1}^{J}(\theta_j - \mu)^2} \\ &< Js^2 \\ &= MSE_{\hat{\theta}}. \end{aligned} \tag{24}$$

See Chapter 1 of Efron (2012) for a simple proof.

This result shows that the unbiased estimator $\hat{\theta}_j$ is inadmissible under squared error loss since it is dominated by $\hat{\theta}_j^{JS}$. This may seem counterintuitive since $\hat{\theta}_j$ is the maximum likelihood estimator of θ_j under model (20). Indeed, as discussed in Efron and Morris (1975), the James/Stein result was initially met with surprise and encountered resistance even among statisticians.

Where does the linear shrinkage rule in equation (23) come from? It should be clear based on the preceding discussion that this estimator is the output of an empirical Bayes procedure based on a normal mixing distribution. Suppose we adopt the prior that $\theta_j \sim \mathcal{N}(\mu, \sigma^2)$. Equation (12) implies that if the hyperparameters μ and σ^2 are known, the posterior mean for θ_j is given by $\theta_j^* = \left(\sigma^2/[s^2 + \sigma^2]\right)\hat{\theta}_j + \left(s^2/[s^2 + \sigma^2]\right)\mu$. If μ is known but σ^2 is not, the shrinkage factor $s^2/(s^2 + \sigma^2)$ must be estimated. The quantity $(J-2)s^2/[\sum_{j=1}^{J}(\hat{\theta}_j - \mu)^2]$ provides an unbiased

estimate of this shrinkage factor as long as $J \geq 3$, resulting in an EB posterior mean equal to $\hat{\theta}_j^{JS}$.[5]

While the James/Stein estimator can be derived from an EB approach based on a normal mixing distribution, linear shrinkage reduces MSE whether or not this model is true. In particular, the result in (24) holds for any configuration of θ_j's and any μ. For the purpose of reducing aggregate MSE, the θ_j's need not be drawn from a normal distribution, or even viewed as random effects at all. The unknown parameters also need not be centered at μ, though shrinkage will result in larger MSE improvements if the prior mean is near the average of the θ_j's. The sum of squares in the denominator of the shrinkage coefficients in equation (23) provides an empirical measure of how far the parameters tend to fall from μ, letting the data dictate the appropriate amount of shrinkage.

Versions of the James/Stein result generalize to more complex settings than the simple homoskedastic noise model in equation (20). These include scenarios where the noise variance s^2 is unknown so must be estimated from the microdata, different noise variances or correlation in noise across units, or an estimated prior mean (Lindley, 1962; Efron and Morris, 1973b; Bock, 1975). The case where the standard error s_j varies across groups is especially relevant since this is likely to be true in any realistic value-added analysis. With at least three units for each value of s_j the linear shrinkage formula in (23) can in principle be implemented separately for every s_j, in which case the basic James/Stein Theorem immediately applies. This amounts to using an unrestricted conditional prior distribution of the form

$$\theta_j | s_j \sim \mathcal{N}(\mu(s_j), \sigma^2(s_j)) \tag{25}$$

for EB shrinkage. In practice shrinking separately for each value of s_j may not be feasible, which motivates less flexible priors imposing restrictions on how the distribution of θ_j varies with s_j. An important caution is that such restrictions may break the James/Stein guarantee of a reduction in MSE.[6] The standard EB posterior mean in (15) assumes independence of θ_j and s_j, and may perform poorly if effect sizes and standard errors are correlated (Chen, 2023). I return to such precision-dependence issues in Section 2.6.

Compound decision problems

The James/Stein Theorem shows that EB shrinkage reduces MSE even when the normal random effects model motivating this method is wrong. Then why does shrinkage yield improvements? The requirement that $J \geq 3$ in the James/Stein Theorem provides a clue. Rather than an assumption on the data generating process, the value of EB shrinkage depends on the objective of the econometric

[5] With normal noise and a normal mixing distribution, the marginal distribution of each estimate is $\hat{\theta}_j \sim \mathcal{N}\left(\mu, s^2 + \sigma^2\right)$. This implies the scaled sum of squared deviations $(s^2 + \sigma^2)^{-1} \sum_{j=1}^{J} (\hat{\theta}_j - \mu)^2$ follows a χ^2 distribution with J degrees of freedom. Its reciprocal then follows an inverse χ^2 distribution, which has mean $1/(J-2)$ for $J \geq 3$. It follows that $E[(J-2)s^2 / \sum_{j=1}^{J} (\hat{\theta}_j - \mu)^2] = s^2/(s^2 + \sigma^2)$ when $J \geq 3$.

[6] Since the James/Stein Theorem implies a reduction in MSE for shrinkage toward any constant, it is not necessary to allow the prior mean to depend on s_j to guarantee an improvement. Implementing the shrinkage formula in (23) separately for each s_j – but shrinking all units towards a common prior mean μ – reduces aggregate MSE as long as there are at least 3 units with each value of s_j. In contrast, imposing restrictions on how the prior variance varies with s_j may break the James/Stein guarantee.

analyst. The gains highlighted in the James/Stein Theorem arise because of the form of the loss function in equation (21), which cares about total mean squared error summed over all J units. This idea is generalized with a loss function of the form

$$\mathcal{L}(\delta_1, ..., \delta_J; \theta_1, ..., \theta_J) = \sum_{j=1}^{J} \ell(\delta_j, \theta_j), \tag{26}$$

where δ_j represents a decision for unit j and $\ell(\theta_j, \delta_j)$ gives the loss associated with making decision δ_j for a unit with parameter θ_j.

Such a loss function gives rise to a *compound decision problem* (Robbins, 1951; Gu and Koenker, 2016) in which a sequence of similar independent decisions are considered for each of several units and the decisionmaker seeks to minimize aggregate loss. A decisionmaker with this loss function views units as *exchangeable* – the labels on the units are irrelevant, so the decisionmaker does not distinguish between different configurations of outcomes across units that lead to the same total loss. This notion of exchangeability motivates the abstraction in model (6) treating the θ_j's as draws from a common distribution G. The James/Stein Theorem reveals that using this conceptual device to incorporate distributional information via EB shrinkage improves expected outcomes in a compound decision problem.

Applying shrinkage outside the compound decision context can quickly lead to *reductio ad absurdum* arguments against EB methods. Should a labor economist shrink estimates of the elasticity of labor supply, the return to schooling, and the intergenerational income elasticity toward one another?[7] Shrinkage seems inappropriate in this scenario because it is hard to imagine a decision for which total error across these three parameters is the relevant criterion. On the other hand, a framework that seeks to minimize aggregate error seems more natural when studying distributions of parameters across large sets of similar units like firms, schools, or neighborhoods. This makes the EB framework appealing for value-added analysis. Sections 2.5 and 3.6 present further analysis of EB methods in a compound decision setup.

2.3 Practical Shrinkage Issues

This subsection considers some issues of interpretation and implementation that often arise in value-added studies using empirical Bayes methods.

Distributions of true parameters, unbiased estimates, and posterior means

In any EB value-added analysis it is worth keeping in mind the distinction between true unobserved effects, unbiased estimates, and shrunk posterior means. For instance, suppose we are interested in summarizing heterogeneity in effectiveness across schools by looking at the variance of school value-added. This is a question about the variability of the true parameters θ_j, which are unobserved.

[7] This facetious example intentionally mimics Robbins' (1951) discussion of jointly shrinking observations on "a butterfly in Ecuador...an oyster in Maryland...[and] the temperature of a star." As noted by Lai and Siegmund (2018), examples of this type led to early skepticism regarding the practical value of the compound decision framework (see also Efron and Morris, 1973a).

Due to sampling error, the variance of the observed unbiased estimates $\hat{\theta}_j$ is too big relative to the variance of true effects:

$$Var(\hat{\theta}_j|s_j) = \sigma_\theta^2 + s_j^2 > \sigma_\theta^2. \tag{27}$$

In contrast, the variance of the oracle posterior means θ_j^* is too small relative to the variance of true effects:

$$Var(\theta_j^*|s_j) = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right)\sigma_\theta^2 < \sigma_\theta^2. \tag{28}$$

With precisely-estimated hyperparameters the EB posterior means $\hat{\theta}_j^*$ should be expected to exhibit variance less than σ_θ^2 as well. As a result, the mixing variance σ_θ^2 will typically lie in between the variances of $\hat{\theta}_j$ and $\hat{\theta}_j^*$ across groups.

The bias-corrected variance estimator in equation (9) splits this difference to recover the variance of unobserved effects. This estimator removes the expected contribution of noise from the variance of $\hat{\theta}_j$'s, yielding a consistent estimate of the true mixing variance σ_θ^2. The estimate $\hat{\sigma}_\theta^2$ is therefore the right metric for summarizing variation in effects across units. More generally, the deconvolution estimate $\hat{G}$ from step 2 of the EB recipe is the appropriate object to use for studying questions about the distribution of latent parameters. A plot of the shrunk posterior means $\hat{\theta}_j^*$ should not be expected to reproduce features of the mixing distribution G.

Shrinkage and regression

Another common application of EB shrinkage arises in the context of group-level OLS regression. Suppose we are interested in an OLS regression of an observed variable for unit j, Z_j, on the unknown parameter θ_j:

$$Z_j = \alpha_0 + \alpha_1\theta_j + e_j. \tag{29}$$

The slope coefficient $\alpha_1 = Cov(Z_j, \theta_j)/Var(\theta_j)$ measures the change in Z_j associated with a one-unit increase in θ_j. Examples of analyses in this mold include Chetty et al.'s (2014b) study linking labor market outcomes to teacher test score effects and Abdulkadiroglu et al.'s (2020) study relating measures of school popularity to test score value-added.

Regression (29) is not feasible since the regressor θ_j is not observed, and substituting the noisy estimate $\hat{\theta}_j$ in its place biases the resulting slope coefficient. To see this, suppose the standard deviation of noise in $\hat{\theta}_j$ is common across units ($s_j = s\ \forall j$) and the estimation error $\hat{\theta}_j - \theta_j$ is uncorrelated with e_j. Then the slope coefficient from a regression of Z_j on $\hat{\theta}_j$ is

$$\begin{aligned}\frac{Cov(Z_j, \hat{\theta}_j)}{Var(\hat{\theta}_j)} &= \frac{Cov(\alpha_0 + \alpha_1\theta_j + e_j, \theta_j + (\hat{\theta}_j - \theta_j))}{Var(\theta_j + (\hat{\theta}_j - \theta_j))} \\ &= \left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s^2}\right)\alpha_1. \end{aligned} \tag{30}$$

Putting the noisy estimate $\hat{\theta}_j$ on the right-hand side of a regression therefore yields bias toward zero. This is an instance of the standard result that classical measurement error in a regressor causes attenuation bias.

Empirical Bayes shrinkage corrects this bias. The regression of Z_j on the oracle posterior mean θ_j^* yields

$$\begin{aligned}\frac{Cov(Z_j, \theta_j^*)}{Var(\theta_j^*)} &= \frac{Cov\left(Z_j, \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s^2}\right)\hat{\theta}_j + \left(\frac{s^2}{\sigma_\theta^2+s^2}\right)\mu_\theta\right)}{Var\left(\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s^2}\right)\hat{\theta}_j + \left(\frac{s^2}{\sigma_\theta^2+s^2}\right)\mu_\theta\right)} \\ &= \left(\frac{\sigma_\theta^2+s^2}{\sigma_\theta^2}\right)\frac{Cov(Z_j, \hat{\theta}_j)}{Var(\hat{\theta}_j)} \\ &= \alpha_1. \end{aligned} \tag{31}$$

Equation (31) shows that using the shrunk posterior mean θ_j^* as a regressor recovers the same coefficient as using the true θ_j. This is a version of errors-in-variables regression, which uses the signal-to-noise ratio of the independent variable to correct attenuation bias due to measurement error. An empirically-feasible version of this correction substitutes the EB posterior mean $\hat{\theta}_j^*$ in place of θ_j^*.[8] Note that while shrinkage fixes attenuation bias due to noise in $\hat{\theta}_j$, this noise still reduces the precision of the resulting estimate since the shrunk regressor θ_j^* has lower variance than the ideal regressor θ_j by equation (28).

The situation is reversed when the unknown parameter θ_j appears on the left-hand side of a regression rather than the right. Consider the OLS regression

$$\theta_j = \beta_0 + \beta_1 Z_j + u_j. \tag{32}$$

The slope coefficient $\beta_1 = Cov(\theta_j, Z_j)/Var(Z_j)$ measures the change in θ_j associated with a one unit increase in Z_j. For example, Chetty and Hendren (2018) investigate the correlates of neighborhood quality by regressing neighborhood effects on location characteristics, and Kline et al. (2020) study the attributes of high-paying employers by regressing firm earnings effects on firm covariates. Assuming the estimation error $\hat{\theta}_j - \theta_j$ is independent of Z_j across units, replacing θ_j with the unbiased estimate $\hat{\theta}_j$ on the left-hand side of (32) recovers the target regression coefficient:

$$\frac{Cov(\hat{\theta}_j, Z_j)}{Var(Z_j)} = \frac{Cov(\theta_j, Z_j)}{Var(Z_j)} + \frac{Cov(\hat{\theta}_j - \theta_j, Z_j)}{Var(Z_j)} = \beta_1. \tag{33}$$

This is a consequence of the fact that classical measurement error on the left-hand side of a regression does not lead to bias. In contrast, shrinkage on the left introduces non-classical measurement error that does generate bias. The slope coefficient from a regression of θ_j^* on Z_j is

[8] When the standard error s_j varies across units a regression of Z_j on θ_j^* recovers α_1 if θ_j is independent of s_j, but may not recover α_1 when θ_j and s_j are correlated (see Section 2.6). One solution in this heteroskedastic case is to shrink all of the $\hat{\theta}_j$'s using a pooled signal-to-noise ratio equal to $\kappa = \sigma_\theta^2/Var(\hat{\theta}_j)$, which can be estimated with $\hat{\kappa} = \hat{\sigma}_\theta^2/[J^{-1}\sum_j(\hat{\theta}_j - \hat{\mu}_\theta)^2]$. A regression of Z_j on $\kappa\hat{\theta}_j$ recovers α_1 regardless of correlation between θ_j and s_j, though this estimator will be less efficient than regressing Z_j on θ_j^* when θ_j and s_j are independent.

$$\frac{Cov(\theta_j^*, Z_j)}{Var(Z_j)} = \frac{Cov\left(\left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s^2}\right)\hat{\theta}_j + \left(\frac{s^2}{\sigma_\theta^2+s^2}\right)\mu_\theta, Z_j\right)}{Var(Z_j)} = \left(\frac{\sigma_\theta^2}{\sigma_\theta^2+s^2}\right)\beta_1. \tag{34}$$

Shrinkage of the dependent variable therefore attenuates the resulting coefficient toward zero. This combination of results suggests a rule-of-thumb for EB shrinkage in a regression context: shrinkage on the right fixes bias, but shrinkage on the left causes bias.

Long regression or mean residuals? Correlated versus uncorrelated random effects

The initial estimation step in many empirical Bayes value-added analyses is an OLS regression like (5), which includes a set of group indicator variables along with a vector of controls X_i. Some studies instead start by partialing controls out of the outcome variable, then apply empirical Bayes methods to group means of the resulting residuals (e.g., Chetty et al., 2014a; Jackson et al., 2020). How should we interpret differences in results between these approaches?

This question can be answered with the familiar OLS omitted variable bias formula. Consider a mean residual strategy applied to school value-added estimation in Example 2. Let $\gamma^s = E[X_i X_i']^{-1} E[X_i Y_i]$ denote the vector of coefficients from a short regression of Y_i on only the controls X_i, omitting the school indicators D_{ij}. This short regression coefficient is linked to the long regression coefficient γ from equation (5) by the relation

$$\begin{aligned} \gamma^s &= \gamma + E\left[X_i X_i'\right]^{-1} E\left[X_i \left(\textstyle\sum_j D_{ij}\theta_j\right)\right] \\ &= \gamma + E\left[X_i X_i'^{-1}\right] E\left[\bar{X}_{\mathbf{J}(i)}\theta_{\mathbf{J}(i)}\right], \end{aligned} \tag{35}$$

where $\mathbf{J}(i) = \{j : \ D_i = j\}$ encodes the school attended by student i, $\bar{X}_j = E[X_i | D_i = j]$ is the mean of X_i at school j, and the second line uses the law of iterated expectations. The school means of the short regression residuals $Y_i - X_i'\gamma^s$ are then

$$\begin{aligned} \theta_j^s &= E\left[Y_i - X_i'\gamma^s | D_i = j\right] \\ &= E\left[\theta_j - X_i'(\gamma^s - \gamma) | D_i = j\right] \\ &= \theta_j - \bar{X}_j' E\left[X_i X_i'\right]^{-1} E\left[\bar{X}_{\mathbf{J}(i)}\theta_{\mathbf{J}(i)}\right]. \end{aligned} \tag{36}$$

Equation (36) shows that the value-added parameters from the long regression and mean residual approaches coincide when $\bar{X}_j' E[X_i X_i']^{-1} E[\bar{X}_{\mathbf{J}(i)}\theta_{\mathbf{J}(i)}] = 0 \ \forall j$. One scenario satisfying this condition is when the averages of the control variables do not vary across groups, so $\bar{X}_j = E[X_i] = 0 \ \forall j$ (recall that the mean of X_i is normalized to zero). This is unlikely in realistic applications since the motivation for including controls in the first place is typically to adjust for group differences in the covariates. In the school value-added context, this would require every school to enroll a student population with the same demographics and mean past achievement.

A second possible justification for the mean residual approach is an assumption that $E[\bar{X}_{\mathbf{J}(i)}\theta_{\mathbf{J}(i)}] = 0$. This allows for the possibility that the school average covariates $\bar{X}_j$ differ across schools but rules

out a relationship between these averages and school value-added θ_j. When X_i is a lagged student test score, for example, this restriction requires schools enrolling students with higher average past achievement to be no better (or worse) in terms of causal value-added. The mean residual approach is therefore implicitly based on an *uncorrelated random effects* model in which the θ_j's are independent of $\bar{X}_j$. In contrast, the long regression approach is compatible with a *correlated random effects* (CRE) model in which the distribution of θ_j may depend arbitrarily on $\bar{X}_j$ (Chamberlain, 1982).[9] Since there is typically no reason to rule out a relationship between θ_j and $\bar{X}_j$, the long regression approach is likely to be preferable in most applications. The uncorrelated random effects assumption can also be tested with a Hausman (1978) test comparing coefficients from the long and short regressions.[10]

2.4 Generalizations of Linear Shrinkage

This subsection considers extensions of the basic empirical Bayes recipe to cases incorporating observed unit characteristics, multiple parameters per unit, and multiple estimates of each unit's parameter.

Adding covariates

Suppose we observe a vector Z_j of covariates for each unit j (including a constant). Consider a conditional normal/normal model of the form:

$$\hat{\theta}_j|\theta_j, s_j, Z_j \sim \mathcal{N}(\theta_j, s_j^2), \tag{37}$$

$$\theta_j|s_j, Z_j \sim \mathcal{N}\left(Z_j'\mu, \sigma_r^2\right), \tag{38}$$

where μ describes the relationship between effect sizes θ_j and observed characteristics Z_j, and σ_r^2 measures residual variance in θ_j not explained by Z_j. By the logic of Section 2.3, a regression of $\hat{\theta}_j$ on Z_j yields an unbiased estimate $\hat{\mu}$ of μ. We can then estimate the unexplained variance in θ_j by deconvolving the resulting residuals with the estimator $\hat{\sigma}_r^2 = J^{-1}\sum_j[(\hat{\theta}_j - Z_j'\hat{\mu})^2 - s_j^2]$.

The posterior mean for θ_j implied by equations (37) and (38) is:

$$\theta_j^* = \left(\frac{\sigma_r^2}{\sigma_r^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\sigma_r^2 + s_j^2}\right)Z_j'\mu. \tag{39}$$

Plugging $\hat{\mu}$ and $\hat{\sigma}_r^2$ into equation (39) yields an EB posterior mean that shrinks $\hat{\theta}_j$ toward an estimated linear index $Z_j'\hat{\mu}$ rather than toward a constant. This approach borrows more strength from units that are similar on observed dimensions, potentially yielding further reductions in MSE

[9] Note that the long regression control coefficient γ can be obtained either by regressing Y_i on X_i and the D_{ij}, or by regressing Y_i on X_i controlling for $\bar{X}_{\mathbf{J}(i)}$ (Mundlak, 1978). The long regression value-added coefficients θ_j are means of the residual $Y_i - X_i'\gamma$.

[10] This test is sometimes described as comparing "fixed effects" and "random effects" models, terminology that mixes the question of whether the θ_j's are treated as fixed or random with the question of whether they are related to $\bar{X}_j$.

relative to $\hat{\theta}_j$. The resulting EB posterior places less weight on the noisy $\hat{\theta}_j$ if the covariates explain more of the variance in θ_j, reflected in a smaller residual variance σ_r^2. It is straightforward to allow the residual variance σ_r^2 to depend on Z_j as well.

Multivariate EB

Suppose we expand the school value-added setup in Example 2 so that schools may affect each of K outcomes for a given student, labeled $Y_{i1}, ..., Y_{iK}$. These might include test scores as well as longer-term outcomes like educational attainment, criminal justice involvement, or earnings (e.g., Jackson et al., 2020; Beuermann et al., 2022; Rose et al., 2022). This yields a separate version of the VAM regression in equation (5) for each outcome:

$$Y_{ik} = \sum_j \theta_{jk} D_{ij} + X_i' \gamma_k + \epsilon_{ik}, \ k \in \{1, ..., K\}. \tag{40}$$

School j's quality is now described by a $K \times 1$ vector $\theta_j = (\theta_{j1}, \theta_{j2}, ..., \theta_{jK})'$ collecting its value-added on all outcomes. Seemingly Unrelated Regression (SUR) estimation of system (40) yields an estimate $\hat{\theta}_j$ of this vector, along with a $K \times K$ sampling variance matrix V_j for each school. This matrix has squared standard errors s_{jk}^2 for each $\hat{\theta}_{jk}$ along its diagonal and off-diagonal terms measuring sampling covariances between $\hat{\theta}_{jk}$ and $\hat{\theta}_{jm}$ for $m \neq k$. These covariances will generally be non-zero if the underlying student-level outcomes are correlated.

Extending the normal/normal model to the multivariate case, we have:

$$\hat{\theta}_j | \theta_j, V_j \sim \mathcal{N}(\theta_j, V_j), \tag{41}$$

$$\theta_j | V_j \sim \mathcal{N}(\mu_\theta, \Sigma_\theta), \tag{42}$$

where the mixing distribution in (42) is now parameterized by a $K \times 1$ mean μ_θ and $K \times K$ variance matrix Σ_θ. The off-diagonal elements of Σ_θ describe how schools' effects on different outcomes are related. For example, do schools that boost test scores also boost high school graduation? The deconvolution step estimates these hyperparameters with:

$$\hat{\mu}_\theta = J^{-1} \sum_j \hat{\theta}_j, \ \hat{\Sigma}_\theta = J^{-1} \sum_j [(\hat{\theta}_j - \hat{\mu}_\theta)(\hat{\theta}_j - \hat{\mu}_\theta)' - V_j]. \tag{43}$$

As before, the bias-corrected variance matrix estimator $\hat{\Sigma}_\theta$ removes excess variability in the $\hat{\theta}_{jk}$'s due to sampling error, as well as the contribution of noise to the covariances of estimates across outcomes.

Finally, in the shrinkage step, the multivariate oracle posterior mean implied by equations (41) and (42) is a precision-matrix-weighted average of the unbiased estimate $\hat{\theta}_j$ and the prior mean:

$$\theta_j^* = \left(V_j^{-1} + \Sigma_\theta^{-1} \right)^{-1} \left(V_j^{-1} \hat{\theta}_j + \Sigma_\theta^{-1} \mu_\theta \right). \tag{44}$$

A multivariate EB posterior mean $\hat{\theta}_j^*$ plugs hyperparameter estimates from (43) into (44). This estimator borrows strength across multiple outcomes as well as across the ensemble of schools when predicting any one of the outcome-specific value-added parameters θ_{jk}. For instance, if effects on test scores (outcome 1) and high school graduation (outcome 2) are highly correlated across schools, then the posterior mean graduation effect $\hat{\theta}_{j2}^*$ may place substantial weight on the estimated test score effect $\hat{\theta}_{j1}$, especially if $\hat{\theta}_{j2}$ is much noisier than $\hat{\theta}_{j1}$. Other applications of multivariate EB include value-added models with heterogeneous effects, so that each unit is associated with treatment effect parameters for multiple subgroups (e.g., Abdulkadiroğlu et al., 2020; Avivi, 2024); and instrumental variables settings with imperfect compliance, with each unit characterized by multiple parameters governing outcomes and selection into treatment (e.g., Raudenbush et al., 2012; Walters, 2015).

Combining estimators

The empirical Bayes approach extends naturally to cases with multiple estimates of each parameter, some possibly biased. In the school value-added example, consider a population OLS regression of Y_i on school indicators with controls for X_i:

$$Y_i = \sum_{j=1}^{J} \alpha_j D_{ij} + X_i'\Gamma + \upsilon_i. \tag{45}$$

This equation differs from the causal value-added model (5) because the OLS residual υ_i satisfies $E[D_{ij}\upsilon_i] = 0 \; \forall j$ by definition, while the error term ϵ_i from the causal model may be correlated with the D_{ij}'s if selection-on-observables fails to hold. This possibility is represented by a list of bias parameters $b_j = \theta_j - \alpha_j$ measuring deviations between causal value-added parameters and OLS regression coefficients.

Let $\{\hat{\alpha}_j\}_{j=1}^J$ denote OLS value-added estimates based on fitting equation (45) in a random sample of students. Suppose we also have access to a second set of estimates $\{\hat{\theta}_j\}_{j=1}^J$ from an alternative research design that is not contaminated by selection bias. For example, the $\hat{\theta}_j$'s may be instrumental variables (IV) estimates derived from randomized school entrance lotteries. The $\hat{\theta}_j$'s are assumed to be (asymptotically) unbiased estimates of the causal θ_j's, but are likely to be noisier than the OLS $\hat{\alpha}_j$ estimates. We then have the model:

$$(\hat{\alpha}_j, \hat{\theta}_j)' | \theta_j, b_j, s_{j\alpha}, s_{j\theta}, c_j \sim \mathcal{N}\left((\theta_j + b_j, \theta_j)', \begin{bmatrix} s_{j\alpha}^2 & c_j \\ c_j & s_{j\theta}^2 \end{bmatrix} \right). \tag{46}$$

Here $s_{j\alpha}^2$ and $s_{j\theta}^2$ are the sampling variances of the OLS and IV estimates, and c_j is their sampling covariance. A Hausman (1978)-style overidentification test comparing OLS and IV is a test of the null hypothesis that $b_j = 0 \; \forall j$ (Angrist et al., 2016).[11]

[11] Under the classical Gauss-Markov assumptions and with no selection bias OLS is the efficient linear estimator of the θ_j's, in which case $c_j = s_{j\alpha}^2$ so that $Var(\hat{\theta}_j - \hat{\alpha}_j) = s_{j\theta}^2 - s_{j\alpha}^2$.

If this overidentification test rejects, we may conclude that the OLS estimates are contaminated by selection bias. The biased $\hat{\alpha}_j$'s may be more precise than the unbiased $\hat{\theta}_j$'s, however, so throwing away OLS completely may discard useful information about school quality. An empirical Bayes approach incorporates this information by forming a best guess of value-added for each school that trades off the bias of OLS against the imprecision of IV.

Consider a model of the causal-valued parameters θ_j and bias parameters b_j as draws from a multivariate normal mixing distribution:

$$(\theta_j, b_j)' | s_{j\alpha}, s_{j\theta}, c_j \sim \mathcal{N}\left((\mu_\theta, \mu_b)', \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta b} \\ \sigma_{\theta b} & \sigma_b^2 \end{bmatrix} \right). \tag{47}$$

The variance parameters σ_θ^2 and σ_b^2 describe the variability of causal value-added and selection bias across schools, while the covariance $\sigma_{\theta b}$ determines whether OLS tends to over-rate or under-rate higher value-added schools. These hyperparameters can be estimated by deconvolving the joint distribution of $(\hat{\alpha}_j, \hat{\theta}_j)$. For example, we can estimate the variance of bias across schools as $\hat{\sigma}_b^2 = J^{-1} \sum_j [(\hat{\alpha}_j - \hat{\theta}_j - \hat{\mu}_b)^2 - (s_{j\alpha}^2 + s_{j\theta}^2 - 2c_j)]$, where $\hat{\mu}_b = J^{-1} \sum_j (\hat{\alpha}_j - \hat{\theta}_j)$. The estimator $\hat{\sigma}_b^2$ looks for overdispersion in the difference between OLS and IV beyond what should be expected from sampling error, implying variation in selection bias.

Posterior means $\theta_j^* = E[\theta_j | \hat{\theta}_j, \hat{\alpha}_j, s_{j\alpha}, s_{j\theta}, c_j]$ based on model (46)-(47) provide minimum MSE forecasts of value-added for each school using all available information. In a scenario where the OLS estimates are extremely precise ($s_{j\alpha} \approx 0$), the oracle posterior mean for θ_j is given by:

$$\theta_j^* = \left(\frac{(1-R^2)\sigma_\theta^2}{(1-R^2)\sigma_\theta^2 + s_{j\theta}^2} \right) \hat{\theta}_j + \left(\frac{s_{j\theta}^2}{(1-R^2)\sigma_\theta^2 + s_{j\theta}^2} \right) (\rho(\hat{\alpha}_j - \mu_b) + (1-\rho)\mu_\theta), \tag{48}$$

where $\rho = Cov(\hat{\alpha}_j, \theta_j)/Var(\hat{\alpha}_j) = (\sigma_\theta^2 + \sigma_{\theta b})/(\sigma_\theta^2 + \sigma_b^2 + 2\sigma_{\theta b})$ is the slope coefficient from a regression of θ_j on $\hat{\alpha}_j$ (the *reliability* of OLS), and R^2 is the R-squared from this regression. This linear shrinkage formula can be seen as a version of equation (39) treating the OLS estimate $\hat{\alpha}_j$ as a covariate that predicts θ_j. An empirical implementation plugs deconvolution estimates of ρ, R^2, and other hyperparameters into (48) to produce an EB posterior mean. Chetty and Hendren (2018) use such a strategy to combine quasi-experimental estimates of neighborhood effects based on cross-neighborhood moves with observational estimates based on levels of permanent resident outcomes. Angrist et al. (2017; 2024a) generalize this approach to estimate school value-added in an underidentified scenario where lotteries are unavailable for some schools.

2.5 EB Decision Rules

We have seen that linear shrinkage delivers posterior mean estimates with low mean squared error. It is often of interest to consider goals other than minimizing MSE. In the school value-added example, suppose a school district administrator seeks to select schools with value-added below a cutoff c for reform or closure, and incurs different costs for erroneously selecting high-performing schools (type I errors) or failing to select low-performing schools (type II errors). The objective of the ad-

ministrator is formalized in a compound decision problem of the form of (26), with component-wise loss function

$$\ell(\delta_j, \theta_j) = \delta_j 1\{\theta_j \geq c\} + (1 - \delta_j)1\{\theta_j < c\}\xi, \tag{49}$$

where $\delta_j \in \{0, 1\}$ indicates selection of school j, ξ is the cost of a type II error, and the cost of a type I error is normalized to one. Gu and Koenker (2023b) discuss empirical Bayes tail selection problems with loss functions of this type.

Since the θ_j's are unknown, the administrator chooses a decision rule $\delta(\hat{\theta}_j, s_j)$ to minimize *risk* (expected loss). With J schools the risk of a decision rule δ is given by

$$\begin{aligned}\mathcal{R}(\delta) &= \sum_{j=1}^{J} E\left[\ell(\delta(\hat{\theta}_j, s_j), \theta_j)|s_j\right] \\ &= \sum_{j=1}^{J} \int\int \ell\left(\delta(\hat{\theta}, s_j), \theta\right) \frac{1}{s_j}\phi\left(\frac{\hat{\theta} - \theta}{s_j}\right) d\hat{\theta} dG(\theta). \end{aligned} \tag{50}$$

The risk-minimizing decision rule is $\delta^* = \arg\min_{\delta \in \mathcal{D}} \mathcal{R}(\delta)$, where $\mathcal{D}$ is the set of functions mapping $(\hat{\theta}_j, s_j)$ to binary decisions. With the loss function in (49), the optimal decision rule is

$$\delta^*(\hat{\theta}_j, s_j) = 1\left\{\Pr[\theta_j < c|\hat{\theta}_j, s_j] \geq \frac{1}{1 + \xi}\right\}. \tag{51}$$

That is, the administrator selects schools with sufficiently high posterior probability of falling below the cutoff c, where the confidence necessary to make a selection depends on the relative costs of type I and type II errors. With the normal mixing distribution in (7) this decision rule becomes:

$$\delta^*(\hat{\theta}_j, s_j) = 1\left\{\left(\frac{\sigma_\theta^2}{\sigma_\theta^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\sigma_\theta^2 + s_j^2}\right)\mu_\theta + \sqrt{\frac{\sigma_\theta^2 s_j^2}{\sigma_\theta^2 + s_j^2}}\Phi^{-1}\left(\frac{1}{1 + \xi}\right) \leq c\right\}. \tag{52}$$

An EB decision rule $\hat{\delta}^*(\hat{\theta}_j, s_j)$ plugs estimated hyperparameters $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$ into equation (52).

Equation (52) reveals that the solution to this tail selection problem will not generally select schools based on the posterior mean. The optimal decision is a cutoff in the $(1 + \xi)^{-1}$ posterior quantile, which adds an adjustment to the posterior mean based on the posterior standard deviation $\sqrt{\sigma_\theta^2 s_j^2/(\sigma_\theta^2 + s_j^2)}$. This means schools with the same posterior mean θ_j^* will be treated differently based on their standard errors s_j. Whether this adjustment rewards or punishes schools with large standard errors depends on whether $\Phi^{-1}(1/(1 + \xi))$ is positive or negative, which depends in turn on whether type I or type II errors are more costly (note that this function crosses zero at $\xi = 1$). When the administrator is especially concerned about mistakenly selecting high-performing schools (so ξ is small), she gives the benefit of the doubt to schools with poor posterior means but large standard errors. When the administrator is more concerned about failing to select low performers (so ξ is large), she penalizes schools with large s_j since there is a reasonable chance a school with a very noisy estimate belongs to the lower tail even if its posterior mean appears favorable.

This simple example demonstrates that the right posterior prediction to report in the EB shrinkage step depends on the goal of the analysis and associated loss function. A clear statement of the loss function is therefore a central part of any coherent EB shrinkage analysis. Specifically, what is the purpose of reporting unit-specific posterior estimates, and how do we expect them to be used? Reporting posterior means is sensible if the aim is to maximize average outcomes by directing consumers to units with high value-added. If the goal is to find units in the tail of the distribution while limiting the likelihood of mistakes, a posterior quantile may be more appropriate. Ordered lists of EB posterior predictions may lead to scrutiny of the highest- and lowest-ranked performers, a tendency that Gu and Koenker (2023b) term the "league table mentality." As I discuss further in Section 3.5, neither posterior means nor posterior quantiles are generally optimal if the objective is to accurately convey information about relative ranks.

2.6 Precision-dependence

So far I have focused on models assuming independence of the effect sizes θ_j and sampling variances s_j^2 across units. In practice these parameters may be correlated. Potential sources of such precision-dependence can be seen in the normal-means problem of Example 1. Recall that the sampling variance of $\hat{\theta}_j$ in this case equals $s_j^2 = \sigma_j^2/n_j$. A correlation between θ_j and s_j^2 will arise if the effect size θ_j is related to group size n_j, the within-group outcome variance σ_j^2, or both.

There are often economic reasons to expect such correlations. In the teacher value-added context, a large body of research establishes that teaching skill improves with experience (Staiger and Rockoff, 2010). Lower-quality teachers may also exit the teaching profession faster (Goldhaber et al., 2011). Both of these forces suggest more data will be available to estimate value-added for higher-quality teachers, implying a negative correlation between θ_j and s_j^2. Research on hospital value-added shows that higher-quality hospitals tend to have higher market shares and gain market share over time, potentially generating a positive relationship between hospital quality and sample size (Chandra et al., 2016, 2023). In empirical Bayes meta-analyses of the distribution of research findings across studies, precision-dependence may arise because studies are screened on the basis of statistical significance (publication bias) or because researchers concerned with statistical power choose sample sizes with an eye toward expected effect sizes (Sterne and Narbord, 2004).

To understand the implications of precision-dependence, consider the independent normal/normal model defined by equations (1) and (7). In this model the simple hyperparameter estimates in equations (8) and (9) are not efficient when s_j^2 varies across units. Instead, the following precision-weighted averages provide asymptotically efficient estimates of the mean and variance of G:

$$\hat{\mu}_\theta = \sum_{j=1}^{J} \left(\frac{\left[\sigma_\theta^2 + s_j^2\right]^{-1}}{\sum_{k=1}^{J} \left[\sigma_\theta^2 + s_k^2\right]^{-1}} \right) \hat{\theta}_j, \tag{53}$$

$$\hat{\sigma}_\theta^2 = \sum_{j=1}^{J} \left(\frac{\left[\sigma_\theta^2 + s_j^2\right]^{-2}}{\sum_{k=1}^{J} \left[\sigma_\theta^2 + s_k^2\right]^{-2}} \right) \left[(\hat{\theta}_j - \hat{\mu}_\theta)^2 - s_j^2 \right]. \tag{54}$$

The weights in these formulas depend on the unknown mixing variance σ_θ^2.[12] A one-step maximum likelihood approach continuously updates σ_θ^2 and μ_θ to simultaneously estimate the parameters and efficient weights, while a two-step approach substitutes in weights based on a first-step estimate of σ_θ^2 (e.g. the unweighted estimator in (9)). A simpler alternative is to use weights that are proportional to n_j or s_j^{-2}, which captures some of the gains to precision-weighting without the need to estimate the optimal weights.

Under the assumption that θ_j and s_j^2 are independent, these precision-weighting approaches will produce more precise estimates of the hyperparameters of G than the unweighted averages in (8) and (9). When θ_j and s_j^2 are correlated, however, precision-weighting changes the estimand, so the unequally-weighted averages in (53) and (54) will not generally be consistent for μ_θ and σ_θ^2. The choice of whether or not to precision-weight when estimating the hyperparameters of G in the deconvolution step therefore involves a classic robustness/efficiency tradeoff.

Precision-dependence also affects the shrinkage step of the EB recipe. The standard EB linear shrinkage estimator in equation (15) is based on a prior distribution assuming independence between θ_j and s_j. Its performance will degrade if this assumption is false, and the shrinkage estimate may even perform worse than the unshrunk estimate $\hat{\theta}_j$ with some forms of precision-dependence.

This problem can be seen in a simple example. Suppose the true mixing distribution has mean zero and variance proportional to s_j^2, so that $\theta_j|s_j \sim \mathcal{N}(0, s_j^2\sigma^2)$. The oracle posterior mean in this model is $E[\theta_j|\hat{\theta}_j, s_j] = (\sigma^2/[1+\sigma^2])\hat{\theta}_j$, which uses a constant shrinkage factor that does not depend on s_j^2 and therefore preserves the ordering of the raw $\hat{\theta}_j$'s. Because true effects are more variable for units with higher standard errors, we should not shrink noisier estimates more. A conventional linear shrinkage estimator of the form of (15) will instead apply more shrinkage to units with larger s_j^2, changing the ordering of units relative to $\hat{\theta}_j$ (and the oracle posterior mean). This means that if we aim to select units with high average value-added, the unadjusted $\hat{\theta}_j$ performs better than standard linear shrinkage in this case.[13] I next consider strategies for avoiding such problems by relaxing the assumption that effect sizes and standard errors are independent.

Testing and modeling precision-dependence

The assumption that θ_j and s_j are independent is testable. It is convenient to test this assumption with regression-based procedures checking whether the conditional mean and variance of θ_j depend on s_j. If θ_j and s_j are independent we have $E[\hat{\theta}_j|s_j] = \mu_\theta$ and $E[(\hat{\theta}_j - \mu_\theta)^2 - s_j^2|s_j] = \sigma_\theta^2$. Regressions of $\hat{\theta}_j$ and $(\hat{\theta}_j - \hat{\mu}_\theta)^2 - s_j^2$ on functions of s_j should therefore yield coefficients of zero. Such tests provide a practical first check for the importance of precision-dependence.

If dependence is present, we can incorporate it into the deconvolution step by estimating a conditional mixing distribution that depends on s_j. Chen (2023) studies a class of conditional location scale estimators (CLOSE) that allow the mean and variance of the prior to depend on precision. As a simple parametric version of such a strategy, consider the specification:

[12] Equations (1) and (7) imply the marginal variance of $\hat{\theta}_j$ is $\sigma_\theta^2 + s_j^2$, while the marginal variance of $(\hat{\theta}_j - \mu_\theta)^2$ is $2(\sigma_\theta^2 + s_j^2)^2$. The estimators in equations (53) and (54) are therefore inverse-variance-weighted averages.

[13] See Xie et al. (2012) for an analysis of the performance of various linear shrinkage rules in models with heteroskedasticity.

$$\theta_j = \psi_0 + \psi_1 \log s_j + s_j^{\psi_2} r_j, \tag{55}$$

with $r_j|s_j \sim \mathcal{N}(0, \sigma_r^2)$. This model treats s_j as a covariate that shifts the mean and variance of G, and assumes features of the mixing distribution beyond the first two moments do not depend on precision.

Equation (55) implies the moment conditions $E[\hat{\theta}_j|s_j] = \psi_0 + \psi_1 \log s_j$ and $E[(\hat{\theta}_j - \psi_0 - \psi_1 \log s_j)^2 - s_j^2|s_j] = s_j^{2\psi_2}\sigma_r^2$. The deconvolution step leverages these restrictions to estimate the hyperparameters of the conditional mixing distribution. A simple two-step approach is to estimate ψ_0 and ψ_1 with a linear regression of $\hat{\theta}_j$ on $\log s_j$, then estimate ψ_2 and σ_r^2 with a non-linear least squares regression of $[(\hat{\theta}_j - \hat{\psi}_0 - \hat{\psi}_1 \log s_j)^2 - s_j^2]$ on $s_j^{2\psi_2}\sigma_r^2$. Alternatively, we can estimate all the parameters in one step by embedding both moment conditions in a single generalized method of moments (GMM) procedure.

The shrinkage step incorporates precision-dependence by forming EB posteriors for the residuals r_j and transforming these residuals to produce posteriors for θ_j. Let $\hat{r}_j = (\hat{\theta}_j - \hat{\psi}_0 - \hat{\psi}_1 \log s_j)/s_j^{\hat{\psi}_2}$ denote a noisy estimate of r_j constructed using hyperparameter estimates from the deconvolution step. Equations (1) and (55) imply $\hat{r}_j|r_j, s_j \sim \mathcal{N}\left(r_j, s_j^{2(1-\psi_2)}\right)$. This motivates the residual linear shrinkage estimator:

$$\hat{r}_j^* = \left(\frac{\hat{\sigma}_r^2}{\hat{\sigma}_r^2 + s_j^{2(1-\hat{\psi}_2)}}\right)\hat{r}_j. \tag{56}$$

An EB posterior mean for θ_j accounting for precision-dependence is then given by:

$$\hat{\theta}_j^* = \hat{\psi}_0 + \hat{\psi}_1 \log s_j + s_j^{\hat{\psi}_2}\hat{r}_j^*. \tag{57}$$

When effect sizes and precision are correlated, this estimator should be expected to improve aggregate MSE relative to both the unbiased estimate $\hat{\theta}_j$ and the conventional linear shrinkage estimator in (15) which neglects precision dependence.

Variance-stabilizing transformations

In some cases it is possible to transform the estimates $\hat{\theta}_j$ to have constant sampling variance rather than estimating the relationship between θ_j and s_j. Such *variance-stabilizing transformations* (VSTs) eliminate concerns about precision-dependence since effect sizes cannot be correlated with precision if precision is constant. In the normal noise model (1), suppose the squared standard error s_j^2 is a known function of the unobserved effect θ_j:

$$s_j^2 = v(\theta_j). \tag{58}$$

Consider a transformation of the form:

$$t(\theta) = \eta \int_{-\infty}^{\theta} v(u)^{-1/2} du \tag{59}$$

for a constant η. The delta method implies

$$\begin{aligned} Var(t(\hat{\theta}_j)|\theta_j) &\approx t'(\theta_j)^2 s_j^2 \\ &= \eta \left(v(\theta_j)^{-1/2} \right)^2 v(\theta_j) \\ &= \eta, \end{aligned} \tag{60}$$

where the second line follows from Leibniz's rule. The transformed variable $\hat{t}_j = t(\hat{\theta}_j)$ then has the same sampling variance for all units, and its mixing distribution can be recovered by deconvolution without worrying about precision-dependence.

A VST is an attractive option for handling precision-dependence in settings where a known function $v(\cdot)$ satisfying (58) is available. This approach is less appealing when the right VST is unknown, though in some cases it may be possible to estimate $v(\cdot)$. This involves modeling s_j^2 as a deterministic function of θ_j as in equation (58), rather than modeling θ_j as a (random) function of s_j as in equation (55).[14]

Example 3: Binomial mixtures

Suppose the microdata Y_i are generated by independent Bernoulli trials with a different success probability θ_j for each group j. This implies the group-specific success counts $C_j = \sum_i D_{ij} Y_i$ follow a mixture of Binomial distributions:

$$C_j|\theta_j, n_j \sim Bin(n_j, \theta_j). \tag{61}$$

The variance of the binomial distribution is given by $Var(C_j|\theta_j, n_j) = n_j\theta_j(1-\theta_j)$. The variance of the success rate $\hat{\theta}_j = n_j^{-1} C_j$ is then deterministically linked to the success probability as $s_j^2 = n_j^{-1}\theta_j(1-\theta_j)$.

With binomial noise, sampling variance is stabilized by the Bartlett (1936) arcsine-square-root transformation:

$$t(\theta) = \sin^{-1}\left(\sqrt{\theta}\right). \tag{62}$$

The derivative of this transformation is $t'(\theta) = \left(2\sqrt{\theta(1-\theta)}\right)^{-1}$, so the asymptotic variance of

[14] Holland (1973) shows that a VST may not exist in the multivariate case where θ_j is a vector with more than one element.

$\hat{t}_j = \sin^{-1}\left(\sqrt{\hat{\theta}_j}\right)$ is given by

$$\begin{aligned} Var(\hat{t}_j|\theta_j, n_j) &\approx \left(\frac{1}{2\sqrt{\theta_j(1-\theta_j)}}\right)^2 \left(\frac{\theta_j(1-\theta_j)}{n_j}\right) \\ &= \frac{1}{4n_j}. \end{aligned} \tag{63}$$

The variance of $\hat{t}_j$ no longer depends directly on θ_j, though some precision-dependence may remain if the sample sizes n_j vary and are correlated with θ_j. Kline et al. (2024) apply the arcsine-square-root transformation to analyze variation in callback rates across first names in a correspondence experiment. Brown (2008) considers an extended class of related binomial VSTs with improved finite-sample properties.

Noisy standard errors

While the standard errors s_j^2 are often treated as known, in practice it is usually necessary to estimate the sampling variance of each $\hat{\theta}_j$. Recall from Example 1 that with normal microdata the sample mean $\hat{\theta}_j$ is normally distributed with mean θ_j and variance $s_j^2 = \sigma_j^2/n_j$. If σ_j^2 is unknown it must be estimated from the data, typically with the unbiased variance estimator:

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1}\sum_i D_{ij}(Y_i - \hat{\theta}_j)^2. \tag{64}$$

The standard error estimate for unit j is then $\hat{s}_j^2 = \hat{\sigma}_j^2/n_j$, which includes error due to noise in $\hat{\sigma}_j^2$.

In principle we can account for this noise using a model for the sampling distribution of $\hat{s}_j^2$. When the microdata Y_i are normally distributed, for example, the scaled sum of squares $(n_j - 1)\hat{\sigma}_j^2/\sigma_j^2$ follows a χ^2 distribution with $n_j - 1$ degrees of freedom. This implies $\hat{s}_j^2$ follows a Gamma distribution with shape parameter $(n_j - 1)/2$ and scale parameter $2\sigma_j^2/[n_j(n_j - 1)] = 2s_j^2/[n_j - 1]$:

$$\hat{s}_j^2|s_j^2, n_j \sim \Gamma\left(\frac{n_j - 1}{2}, \frac{2s_j^2}{n_j - 1}\right). \tag{65}$$

This model can be used to implement a bivariate deconvolution that estimates the joint distribution of (θ_j, s_j^2) accounting for the noise in both $\hat{\theta}_j$ and $\hat{s}_j^2$ (see, e.g., Gu and Koenker, 2017, 2023a). Multivariate deconvolution may be empirically challenging, however, especially if precision-dependence is present so that the effect sizes θ_j and group-specific variances σ_j are correlated with one another or with sample size n_j.

Three considerations suggest the sampling error in $\hat{s}_j$ may not be consequential in typical applications. First, noise in estimates of sampling variance does not compromise estimates of some features of the mixing distribution. For example, the bias-corrected variance estimate in (9) is consistent for σ_θ^2 whether the true noise variance s_j^2 or an unbiased estimate $\hat{s}_j^2$ is used for bias correction (see Section 3.1 for elaboration on this point). Second, the noise in $\hat{s}_j^2$ will often be

approximately independent of the noise in $\hat{\theta}_j$, ruling out more pernicious forms of measurement error in $\hat{s}_j^2$. With normal data the sample mean and sample variance are independent.[15] Finally, the noise in $\hat{s}_j^2$ disappears with sample size at a faster rate than the noise in $\hat{\theta}_j$. Equation (65) implies $\hat{s}_j^2$ has variance $2s_j^4/(n_j-1)$, which will typically be much smaller than s_j^2 itself as long as n_j is reasonably large. This motivates the standard approach of treating the noise in $\hat{s}_j^2$ as negligible relative to the noise in $\hat{\theta}_j$.

2.7 Connections to Machine Learning

Empirical Bayes methods are closely connected to machine learning (ML) approaches to model selection and regularization. ML refers to a battery of empirical tools used to reduce overfitting in high-dimensional settings where the number of available predictors is large relative to sample size (see Varian, 2014, Mullainathan and Spiess, 2017, and Athey and Imbens, 2019 for econometrically-oriented reviews). The idea of penalizing model complexity to limit overfitting bears close resemblance to the use of EB shrinkage to reduce variance. I next develop this connection further by highlighting well-known equivalences between EB shrinkage and simple machine learning approaches.

Return to the setup of Example 1 with normally-distributed microdata, and simplify further by assuming a common standard deviation of outcomes across groups ($\sigma_j = \sigma_y \ \forall j$). Suppose the θ_j's are drawn from the normal mixing distribution in (7) and set the prior mean μ_θ to zero. The oracle posterior density for θ_j can be written

$$p(\theta_j|\mathbf{D},\mathbf{Y}) = \frac{\left[\prod_{i=1}^{N}\left\{\frac{1}{\sigma_y}\phi\left(\frac{Y_i-\theta_j}{\sigma_y}\right)\right\}^{D_{ij}}\right]\frac{1}{\sigma_\theta}\phi\left(\frac{\theta_j}{\sigma_\theta}\right)}{\int\left[\prod_{i=1}^{N}\left\{\frac{1}{\sigma_y}\phi\left(\frac{Y_{ij}-\theta}{\sigma_y}\right)\right\}^{D_{ij}}\right]\frac{1}{\sigma_\theta}\phi\left(\frac{\theta}{\sigma_\theta}\right)d\theta}, \tag{66}$$

where $\mathbf{D}$ and $\mathbf{Y}$ are vectors collecting the group indicators D_{ij} and outcomes Y_i for all observations. We have already seen by equation (11) that this posterior distribution is normal with mean given by the linear shrinkage formula θ_j^* in (12). Since the mean and mode of a normal distribution coincide, we can equivalently represent θ_j^* as the maximizer of the posterior density in (66), which implies

$$\begin{aligned}(\theta_1^*,...,\theta_J^*) &= \arg\max_{\theta_1,...,\theta_J}\sum_{j=1}^{J}\log p(\theta_j|\mathbf{D},\mathbf{Y}) \\ &= \arg\max_{\theta_1,...,\theta_J}\sum_{j=1}^{J}\sum_{i=1}^{N}D_{ij}\log\phi\left(\frac{Y_{ij}-\theta_j}{\sigma_y}\right)+\sum_{j=1}^{J}\log\phi\left(\frac{\theta_j}{\sigma_\theta}\right),\end{aligned} \tag{67}$$

[15] Independence of $\hat{\theta}_j$ and $\hat{s}_j^2$ does not generally hold for estimators other than the sample mean. For instance, Keane and Neal (2023) note that instrumental variables estimation produces correlated noise in estimates and standard errors.

where I have dropped constants that do not depend on the θ_j's. The maximizer of the posterior density is called a *maximum a posteriori* (MAP) estimate. Plugging normal densities into equation (67) and simplifying yields

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

$$= \arg\min_{\theta_1,...,\theta_J} \sum_{j=1}^{J}\sum_{i=1}^{N} D_{ij}(Y_{ij} - \theta_j)^2 + \lambda h(\theta_1, ..., \theta_J), \tag{68}$$

where $\lambda = \sigma_y^2/\sigma_\theta^2$ and $h(\theta_1, ..., \theta_J) = \sum_j \theta_j^2$.

Equation (68) shows that the linear shrinkage posteriors θ_j^* solve a regularized least squares problem with an L2 (quadratic) penalty function $h(\cdot)$, a simple ML procedure known as *ridge regression*. This implies we can equivalently think of ridge regression as a Bayesian procedure based on an independent mean-zero normal prior over the parameters. An EB version of this procedure plugs in an estimate of the prior, which means using the data to choose λ. The EB posterior means $\hat{\theta}_j^*$ are therefore ridge regression estimates based on a data-dependent value of the ridge penalty.

ML regularization procedures can often be reinterpreted through an EB lens, with the specific form of regularization determined by choices of prior distribution and loss function. This EB view helps to clarify the implicit distributional assumptions and objectives underlying ML procedures. For example, if we choose a Laplace (double-exponential) prior distribution rather than a normal distribution for G, an analogous derivation yields a MAP estimator that replaces the L2 penalty in equation (68) with an L1 (absolute value) penalty $h(\theta_1, ..., \theta_J) = \sum_j |\theta_j|$. Regularized least squares with an L1 penalty is another basic ML procedure called the *least absolute shrinkage and selection operator* (lasso; Tibshirani, 1996). The use of lasso can therefore be justified by a choice of Laplace prior combined with a choice to report MAP estimates rather than posterior means.[16] Abadie and Kasy (2019) consider the risk properties of ridge, lasso, and other common ML procedures in an EB framework.

2.8 Linear Shrinkage Application: School Value-Added in Boston

The basic empirical Bayes recipe is illustrated here by estimating school value-added in Boston based on data from Angrist et al. (2017). I focus on 2014 math value-added estimates for 46 Boston middle schools. Thirty of these 46 schools operate within the Boston Public Schools (BPS) district, while the remaining 16 are charter schools, which are publicly-funded schools that operate outside BPS and have more freedom than traditional public schools to set curricula and make staffing decisions. The charters in this sample mostly follow the "No Excuses" educational model, a package of practices that has been shown to generate large achievement gains for students in Boston and elsewhere (Abdulkadiroğlu et al., 2011; Angrist et al., 2013; Dobbie and Fryer, 2013;

[16] Tibshirani (1996) noted this Bayesian interpretation when first introducing the lasso.

Walters, 2018). Further details on the characteristics of Boston schools and students are available in Angrist et al. (2017).

Figure 1 summarizes the three steps of the empirical Bayes recipe for Boston middle schools. I implement step 1 with an OLS regression of sixth-grade math test scores on school indicators and controls as in equation (5). Math scores Y_i are standardized to have mean zero and standard deviation (σ) one in the Boston student population. The covariate vector X_i includes fifth-grade math and reading scores along with indicators for sex, race, free or reduced price lunch status, special education, and English language learner status. I center the school coefficient estimates to have mean zero so that each $\hat{\theta}_j$ is interpretable as an estimate of the effect of school j relative to the average school. Standard errors s_j are computed with the White (1980) heteroskedasticity-robust variance estimator. The open bars in panel A of Figure 1 display a histogram of the estimated $\hat{\theta}_j$'s, with blue bars representing traditional BPS schools and red bars representing charter schools. These value-added estimates range from roughly -0.4σ to 0.5σ with a standard deviation of 0.221σ.

Some of the variation in OLS VAM estimates in Figure 1 comes from statistical noise in the estimated $\hat{\theta}_j$'s. Step 2 of the empirical Bayes recipe adjusts for this noise to recover an estimate of the underlying distribution of school quality. The average s_j^2 in this sample is 0.010, which (using equation (9)) results in a bias-corrected standard deviation estimate equal to $\hat{\sigma}_\theta = \sqrt{0.221^2 - 0.010} = 0.197\sigma$. The black curve in panel A of Figure 1 plots a normal distribution with standard deviation $\hat{\sigma}_\theta$, which is the deconvolution estimate of G under the normal model for the mixing distribution in (7). The raw standard deviation of average test scores across Boston schools is 0.5σ, so the estimated value-added distribution implies that only about $(0.2/0.5)^2 \times 100 = 16\%$ of the observed variance in school performance is due to causal contributions of schools, with the remaining 84% explained by selection bias.

The standard errors s_j vary across schools, so I next probe for dependence between effect sizes and precision as discussed in Section 2.6. This investigation suggests little relationship between school value-added and sampling variance. A regression of $\hat{\theta}_j$ on $\log s_j$ yields a slope coefficient of 0.246 with a robust standard error of 0.231, while a regression of $\hat{\theta}_j^2 - s_j^2$ on $\log s_j$ produces a slope coefficient of 0.012 with a standard error 0.047. In view of these weak relationships between estimates and standard errors, I maintain an independent prior in step 3 of the EB recipe and form linear shrinkage posterior means based on equation (15) for each school.

The resulting linear shrinkage estimates are plotted in the solid histogram in panel A of Figure 1. As expected based on the discussion in Section 2.3, the standard deviation of shrunk posteriors (0.180σ) is smaller than the standard deviation of the prior distribution (0.197σ), which is in turn smaller than the standard deviation of the unadjusted VAM estimates (0.221σ). To understand the gains from shrinkage, note that MSE for $\hat{\theta}_j$ equals the average squared standard error (0.010), while MSE for the posterior mean equals the difference between the variance of the mixing distribution and the variance of posterior means.[17] This comparison indicates that shrinkage reduces aggregate MSE for school value-added estimates by $(1 - [(0.197^2 - 0.180^2)/0.010]) \times 100 = 36\%$.

[17] By the law of total variance we have $E[(\theta_j - E[\theta_j|\hat{\theta}_j, s_j])^2] = Var(\theta_j) - Var(E[\theta_j|\hat{\theta}_j, s_j])$.

The upper tail of the estimated value-added distribution in panel A of Figure 1 is disproportionately composed of charter schools. This suggests that incorporating differences in effectiveness across school sectors may result in improved estimates of school quality. I account for school sector with a conditional prior of the form of (38), allowing a charter-sector location shift in the mean of the mixing distribution and assuming a common within-sector variance σ_r^2.

Following the logic of Section 2.3, I estimate the charter sector effect by regressing the unbiased $\hat{\theta}_j$'s on a charter indicator, which yields a coefficient of 0.293σ. This implies that attending an average charter school boosts achievement by nearly one-third of a standard deviation relative to an average BPS school. The standard deviation of residuals from this regression is 0.194σ, which generates a deconvolution estimate of the residual variance in school quality equal to $\hat{\sigma}_r = \sqrt{0.194^2 - 0.010} = 0.139\sigma$. The smaller value of $\hat{\sigma}_r$ relative to the unconditional $\hat{\sigma}_\theta$ reveals that charter status explains half of the variation in effectiveness across Boston schools: the implied R-squared from a regression of unobserved school quality θ_j on a charter indicator is $1 - (0.139/0.197)^2 = 0.502$. Linear shrinkage posteriors incorporating charter status shrink the $\hat{\theta}_j$'s toward a sector-specific mean using equation (39).

Panel B of Figure 1 summarizes this conditional EB procedure by plotting the histogram of unadjusted VAM estimates, separate estimated prior distributions for BPS and charter schools, and a histogram of the resulting EB posteriors. A comparison to panel A illustrates how conditional shrinkage treats schools differently based on charter sector status. In panel A, for example, shrinkage toward the mean has little effect on charter schools with unbiased $\hat{\theta}_j$'s near zero since these schools are already estimated to be about average. The shrinkage in panel B accounts for the fact that such schools come from a high-performing sector and pulls estimates for these schools up above zero. Incorporating charter status into the shrinkage procedure yields further improvements in mean squared error, reducing aggregate MSE by an estimated 47% relative to the unbiased $\hat{\theta}_j$'s.

3 Non-Parametric Empirical Bayes

This section extends the EB framework by relaxing the normality and independence assumptions maintained for much of Section 2. I first discuss methods for estimating the variance of parameters across units under minimal assumptions, building on the variance component estimation framework of Kline et al. (2020). I then proceed to cover non-parametric estimation of priors and posteriors, partial identification, and other important special cases of non-parametric EB methods. The section concludes with an application of non-parametric EB to a labor market correspondence experiment studied by Kline et al. (2022) and Kline et al. (2024).

3.1 Bias-Corrected Variance Estimation

As before, consider a grouped data structure with J groups, each associated with a group-specific parameter θ_j. Collect these parameters in the $J \times 1$ vector $\Theta = (\theta_1, ..., \theta_J)'$. Suppose we can form an unbiased estimate $\hat{\theta}_j$ of each group's parameter, and collect these estimates in the vector

$\hat{\Theta} = (\hat{\theta}_1, ..., \hat{\theta}_J)'$. A $J \times J$ matrix V describes the sampling variance of $\hat{\Theta}$. This matrix has the squared standard errors s_j^2 of the individual $\hat{\theta}_j$'s along its diagonal. Unlike the simpler framework of Section 2.1, I do not assume the noise in $\hat{\theta}_j$ is normally distributed, and I do not assume the $\hat{\theta}_j$'s are independent (i.e. the off-diagonal elements of V may not be zero). Moreover, I assume the elements of V are unknown, but an unbiased variance estimator $\hat{V}$ is available. This setup is formalized as:

$$E\left[\hat{\Theta}|\Theta, V\right] = \Theta, \tag{69}$$

$$E\left[(\hat{\Theta} - \Theta)(\hat{\Theta} - \Theta)'|\Theta, V\right] = V, \tag{70}$$

$$E\left[\hat{V}|\Theta, V\right] = V. \tag{71}$$

The conditioning on (Θ, V) in (69)-(71) emphasizes that these objects are treated as fixed parameters rather than random variables for now.

Suppose we are interested in a quadratic form in the vector Θ, given by:

$$Q(\Theta, A) = \Theta' A \Theta, \tag{72}$$

where A is a known non-random weighting matrix. An important special case uses the matrix

$$A_0 = \frac{1}{J-1}\left(I_J - \frac{1}{J} 1_J 1_J'\right), \tag{73}$$

where I_J is the $J \times J$ identity matrix and 1_J is a $J \times 1$ vector of 1's. With this choice of weighting matrix, the quadratic form $Q(\Theta, A_0)$ is the sample variance of θ_j's:

$$Q(\Theta, A_0) = \frac{1}{J-1}\sum_{j=1}^{J}\left(\theta_j - \bar{\theta}\right)^2, \tag{74}$$

with $\bar{\theta} = J^{-1}\sum_j \theta_j$. Parameter $Q(\Theta, A_0)$ summarizes variation in the unknown θ_j's across the finite set of J observed groups. The general quadratic form in (72) also nests other useful special cases, such as covariances between value-added parameters across multiple outcomes.[18]

An obvious starting point for estimating $Q(\Theta, A)$ is a plug-in estimator $Q(\hat{\Theta}, A)$, which substitutes the unbiased estimate $\hat{\Theta}$ for the true Θ in equation (72). This estimator is biased due to the

[18] In the multivariate value-added framework of Section 2.4, collect value-added parameters for all schools and outcomes in the $JK \times 1$ vector $\Theta = (\theta_{11}, ..., \theta_{J1}, \theta_{12}, ..., \theta_{JK})'$, and let A_{km} denote a $JK \times JK$ matrix with A_0 at its $k-m$th $J \times J$ block and zeros elsewhere. Then $Q(\Theta, A_{km}) = (J-1)^{-1}\sum_j (\theta_{jk} - \bar{\theta}_k)(\theta_{jm} - \bar{\theta}_m)$ with $\bar{\theta}_k = J^{-1}\sum_j \theta_{jk}$.

non-linearity of the quadratic form $Q(\cdot)$. Specifically, we have

$$\begin{aligned} E\left[Q(\hat{\Theta}, A)|\Theta, V\right] &= E\left[\hat{\Theta}' A \hat{\Theta}|\Theta, V\right] \\ &= \Theta' A \Theta + E\left[\left(\hat{\Theta} - \Theta\right)' A \left(\hat{\Theta} - \Theta\right)|\Theta, V\right] \\ &= Q\left(\Theta, A\right) + tr\left(AV\right), \end{aligned} \tag{75}$$

where the last equality uses standard properties of the trace operator.[19] Equation (75) formalizes the intuition from Section 2.1 that noise inflates the sample variance of estimated $\hat{\theta}_j$'s relative to the variability of the true θ_j's. With the weighting matrix in (73) and a diagonal V, the bias in the plug-in variance estimate is given by $tr(A_0 V) = J^{-1} \sum_j s_j^2$.

The bias in the plug-in estimator can be corrected using the variance estimate $\hat{V}$. The result in equation (75) motivates a bias-corrected estimator of the form:

$$\hat{Q}_{BC} = Q(\hat{\Theta}, A) - tr(A\hat{V}). \tag{76}$$

When $\hat{V}$ is unbiased for V, the second term of (76) is an unbiased estimator of the bias term in (75). This implies $\hat{Q}_{BC}$ is an unbiased estimator of $Q(\Theta, A)$:

$$\begin{aligned} E\left[\hat{Q}_{BC}|\Theta, V\right] &= E\left[Q(\hat{\Theta}, A)|\Theta, V\right] - E\left[tr\left(A\hat{V}\right)|\Theta, V\right] \\ &= Q\left(\Theta, A\right) + tr\left(AV\right) - tr\left(AE[\hat{V}|\Theta, V]\right) \\ &= Q(\Theta, A). \end{aligned} \tag{77}$$

This bias correction approach may be implemented with various choices of the variance estimate $\hat{V}$. The properties of the resulting $\hat{Q}_{BC}$ will depend on this choice as well as the variance structure of the underlying microdata. For the case where $\hat{\Theta}$ is a vector of linear regression coefficients, Andrews et al. (2008) consider bias-correction with a $\hat{V}$ that is unbiased with homoskedasticity but biased and inconsistent with heteroskedasticity. A $\hat{V}$ formed based on conventional White (1980) heteroskedasticity-robust (HC0) standard errors is consistent with heteroskedasticity but biased in finite samples. The HC2 robust variance modification proposed by MacKinnon and White (1985) is unbiased when the microdata are homoskedastic and biased but consistent with heteroskedasticity. Kline et al. (2020) propose a leave-out variance estimator that yields a finite-sample unbiased $\hat{V}$ with arbitrary heteroskedasticity. This property makes the Kline et al. (2020) estimator an appealing choice for bias-correction in value-added studies with few observations per unit. In cases where groups are large enough for the standard errors of the individual $\hat{\theta}_j$'s to be accurate, alternative heteroskedasticity-robust variance matrices are likely to produce similar results.

Unbiased estimation of the mixing variance

Equation (77) establishes that we can use $\hat{\Theta}$ and $\hat{V}$ to construct an unbiased estimate of any

[19] A scalar equals its trace, trace is a linear operator, and a trace is invariant to cyclic permutations.

quadratic form $Q(\Theta, A)$, treating Θ and V as unknown fixed parameters. If we instead view the θ_j's as random effects drawn from a mixing distribution, the law of iterated expectations implies $\hat{Q}_{BC}$ is an unbiased estimate of the expectation of $Q(\Theta, A)$. We have

$$E\left[\hat{Q}_{BC}\right] = E\left[E\left[\hat{Q}_{BC}|\Theta, V\right]\right] = E\left[Q\left(\Theta, A\right)\right], \tag{78}$$

where the outer expectations treat Θ and V as random. Importantly, this implies that in the random effects model (6) with θ_j's drawn independently from G, we can form an unbiased estimate of the mixing variance as:

$$\hat{\sigma}^2_\theta = \hat{\Theta}' A_0 \hat{\Theta} - tr(A_0 \hat{V}). \tag{79}$$

This estimator is unbiased for σ^2_θ because

$$\begin{aligned} E\left[\hat{\Theta}' A_0 \hat{\Theta} - tr(A_0 \hat{V})\right] &= E\left[Q(\Theta, A_0)\right] \\ &= E\left[\frac{1}{J-1}\sum_{j=1}^{J}\left(\theta_j - \bar{\theta}\right)^2\right] \\ &= \sigma^2_\theta, \end{aligned} \tag{80}$$

where the last equality follows from the fact that the sample variance of θ_j's is an unbiased estimate of the population variance. When $\hat{V}$ is a diagonal matrix composed of unbiased squared standard error estimates $\hat{s}^2_j$, the bias-corrected mixing variance estimator simplifies to $\hat{\sigma}^2_\theta = (J-1)^{-1}\sum_j[(\hat{\theta}_j - \hat{\mu}_\theta)^2 - J^{-1}(J-1)\hat{s}^2_j]$. This estimator modifies equation (9) with degrees-of-freedom corrections to yield a finite-sample unbiased estimate of σ^2_θ.

In summary, subtracting off an average squared standard error as in equation (9) is a simple and transparent starting point for bias-corrected mixing variance estimation. Using the more comprehensive bias-correction formula in equation (79) is likely to make a difference when J is small and when the off-diagonal elements of V are large. In a value-added regression like equation (5) with independent student observations, correlation in the $\hat{\theta}_j$'s across schools comes from error in the estimated control coefficient $\hat{\gamma}$.[20] This implies the off-diagonal elements of V can be safely ignored when the control coefficients are estimated very precisely, but may be important in scenarios with many fixed effects or other high-dimensional controls. Other approaches to value-added estimation, such as "movers designs" leveraging individuals switching between groups, may also generate important correlations in noise across units, leading to a non-diagonal V (Bonhomme and Denis, 2024). The full bias correction in (79) extracts the signal variance of value-added from correlated

[20] As noted in Section 2.3, the $\hat{\theta}_j$'s are school means of the long regression residual $Y_i - X_i'\hat{\gamma}$, which are uncorrelated across schools if there is negligible noise in $\hat{\gamma}$. The nature of correlation in estimates will also depend on how the reference category in a value-added regression is defined. Equation (5) includes a dummy for every school and no constant term, which makes $\hat{\theta}_j$ an estimate of an average covariate-adjusted outcome level at school j. If we instead include a constant and omit a dummy for school 1, the resulting value-added estimates for schools 2 through J equal $\hat{\Delta}_j = \hat{\theta}_j - \hat{\theta}_1$, which are likely to be correlated since these are differences relative to the same base school.

noise in such cases. In applications where correlated noise is a potential concern, a comparison of mixing variance estimates based on equations (9) and (79) is a useful robustness check.

Example 4: Worker and firm effects

A prominent application of bias-corrected variance estimation comes from two-way fixed effects (TWFE) models for worker and firm effects on earnings. Starting with Abowd et al. (1999), a large literature studies linear regressions of the form:

$$\log Y_{it} = \alpha_i + \psi_{\mathbf{J}(i,t)} + X_{it}'\beta + \epsilon_{it}, \tag{81}$$

where Y_{it} is earnings for worker i in year t and X_{it} is a vector of time-varying controls. The function $\mathbf{J}(i,t)$ returns the identity of the firm employing worker i in year t. The worker effect α_i represents a permanent component of earnings specific to worker i that is constant across employers, while the firm effect ψ_j reflects an effect of working for firm j that is constant across workers and over time. Under an exogenous mobility assumption the firm effects in equation (81) capture causal effects of particular employers on pay, which may reflect factors such as market power, compensating differentials, wage setting policies, or rent sharing (Card et al., 2018). Analyses of such additive worker and firm effect models include Gruetter and Lalive (2009), Card et al. (2013), Card et al. (2015), Song et al. (2018), Lachowska et al. (2023a), and Lachowska et al. (2023b).

The parameters of equation (81) are identified based on movements of workers across firms, so TWFE estimation must be restricted to a connected set of employers linked by worker mobility. Even within this set there may be few workers linking some groups of firms, yielding noisy estimated worker and firm effects $\hat{\psi}_j$ and $\hat{\alpha}_i$. The errors in these estimates are also likely to be correlated: overestimating a worker's effect will tend to lead to underestimating the effect of his or her firm, resulting in a negative sampling covariance between $\hat{\alpha}_i$ and $\hat{\psi}_{\mathbf{J}(i,t)}$. The influence of this noise on the variation in estimated worker and firm effects is commonly referred to as "limited mobility bias" (Andrews et al., 2008), a special case of the plug-in bias illustrated in equation (75).

The variances and covariances of the worker and firm effects from equation (81) can be written as special cases of the quadratic form in (72). Bonhomme et al. (2023) compare estimates of these variance components using multiple strategies for bias correction including the "homoskedasticity-only" correction of Andrews et al. (2008), the heteroskedasticity-robust bias correction of Kline et al. (2020), and random effects methods building on Woodcock (2008) and Bonhomme et al. (2019). Their results for several countries show that bias-correction substantially reduces the estimated magnitude of variation in firm effects and often flips the sign of the estimated covariance between α_i and $\psi_{\mathbf{J}(i,t)}$ from negative to positive. They also find that alternative methods for bias-correction tend to produce similar results. Adopting an EB view, we can think of the parameters of equation (81) as (correlated) random effects drawn from a mixing distribution describing worker and firm heterogeneity, and interpret bias-corrected variance components as estimates of the second moments of this distribution.

3.2 Non-parametric Priors and Posteriors

The preceding section shows that we can estimate the variance of G under minimal assumptions. If we are content to quantify heterogeneity across units with the mixing variance and form linear shrinkage posteriors with low mean squared error, estimating the mean and variance of G may be enough. In some cases, however, it is useful to develop a more complete picture of the value-added distribution. To flexibly estimate the full mixing distribution G, I return to a hierarchical random effects setup with mutually independent normally distributed estimates and effect sizes independent of standard errors:

$$\hat{\theta}_j|\theta_j, s_j \sim \mathcal{N}(\theta_j, s_j^2), \tag{82}$$

$$\theta_j|s_j \sim G. \tag{83}$$

The mixing distribution is non-parametrically identified in this model (Kotlarski, 1967; Evdokimov and White, 2012), which motivates estimators that avoid imposing a functional form for G. I consider two approaches to flexible estimation of G: a non-parametric maximum likelihood estimator (NPMLE), and a log-spline deconvolution estimator proposed by Efron (2016).

Non-parametric maximum likelihood

Non-parametric maximum likelihood is a classic approach to flexibly estimating distributions of unobserved heterogeneity. NPMLE was outlined in an abstract by Robbins (1950) and developed in detail by Kiefer and Wolfowitz (1956). Within labor economics, Heckman and Singer (1984) applied NPMLE to estimate heterogeneity distributions in mixed proportional hazards models. The NPMLE estimator picks $\hat{G}$ to maximize the likelihood of the observed data over all possible mixing distributions. This estimator is given by:

$$\hat{G}_{NPMLE} = \arg\max_{G \in \mathcal{G}} \sum_{j=1}^{J} \log\left(\int \frac{1}{s_j} \phi\left(\frac{\hat{\theta}_j - \theta}{s_j} \right) dG(\theta) \right), \tag{84}$$

where $\mathcal{G}$ is the set of all cumulative distribution functions. The $\hat{G}_{NPMLE}$ that solves this problem is a discrete distribution with at most J mass points. While maximizing over all possible distribution functions may appear to be a formidable empirical task, Koenker and Mizera (2014) outline computationally-efficient procedures that quickly approximate the NPMLE using convex optimization methods.[21] Gilraine et al. (2020) apply these methods to non-parametrically estimate distributions of teacher value-added.

NPMLE imposes no restrictions on the mixing distribution, and EB decision rules using $\hat{G}_{NPMLE}$ as prior have been shown to provide a close approximation to the decisions of an oracle who knows G (Jiang and Zhang, 2009; Jiang, 2020). However, the discrete distribution function produced by NPMLE can be unweildy in some applications. Posterior distributions using $\hat{G}_{NPMLE}$ as prior

[21] The **REBayes** R package implements these methods (Koenker and Gu, 2017).

will also be discrete, which may be awkward in settings where exact ties of θ_j's are implausible. Koenker (2020) notes that this discreteness can lead to narrow EB credible intervals (differences in posterior quantiles) with poor frequentist coverage.[22] These issues make it attractive to consider adding some smoothness restrictions in non-parametric deconvolution.

Log-spline deconvolution

Efron (2016) proposes a flexible deconvolution approach that approximates the mixing distribution with a smooth log density parameterized by a natural cubic spline. For a grid of M support points $(\bar{\theta}_1, ..., \bar{\theta}_M)$, suppose the probability mass at point m is given by:

$$g_m(\alpha) = \exp\left(S_m'\alpha - \log\left(\sum_{k=1}^{M} \exp(S_k'\alpha)\right)\right), \tag{85}$$

where $S_m = S(\bar{\theta}_m)$ is a $K \times 1$ vector of natural cubic spline basis functions with K knots and α is a $K \times 1$ parameter vector. The parameters of the model are estimated by penalized maximum likelihood as follows:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{j=1}^{J} \log\left(\sum_{m=1}^{M} g_m(\alpha) \frac{1}{s_j} \phi\left(\frac{\hat{\theta}_j - \bar{\theta}_m}{s_j}\right)\right) - \lambda\sqrt{\alpha'\alpha}, \tag{86}$$

where $\lambda \geq 0$ is a tuning parameter. The resulting log-spline deconvolution estimate of G is $\hat{G}_{LS}(\theta) = \sum_m 1\{\bar{\theta}_m \leq \theta\} g_m(\hat{\alpha})$.

The Efron (2016) log spline deconvolution estimator is straightforward to compute and (when M is large) yields a smoother mixing distribution estimate than NPMLE.[23] These advantages make it an appealing option for flexible deconvolution in practice. However, the estimator in (86) requires choosing several tuning parameters including the number of spline knots K; the support limits, spacing, and number of the support points $\bar{\theta}_m$; and the penalization parameter λ. By the logic of Section 2.7, the penalty term $\lambda\sqrt{\alpha'\alpha}$ may be interpreted as a second-level prior that pushes the estimated mixing distribution toward a uniform distribution to manage overfitting. Kline et al. (2022) suggest choosing λ so that the variance of the resulting $\hat{G}_{LS}$ matches a bias-corrected variance estimate based on the methods of Section 3.1. Tuning $\hat{G}_{LS}$ to match low-order moment estimates in this way ensures that the deconvolved mixing distribution reproduces basic features of G estimated with simpler methods (Efron and Tibshirani, 1996).

Non-parametric posteriors

Non-parametric EB shrinkage uses an NPMLE or log-spline estimate $\hat{G}$ when forming EB posterior distributions $\mathcal{P}(\theta|\hat{\theta}_j, s_j; \hat{G})$ for each unit. The resulting distribution can be used to compute any posterior feature of interest. The posterior mean of an oracle who knows G is

[22] Armstrong et al. (2022) propose EB confidence intervals with average coverage guarantees regardless of the form of G.

[23] The **deconvolveR** R package provides software for log spline deconvolution (Narasimhan and Efron, 2020).

$$\theta_j^* = \int \theta d\mathcal{P}\left(\theta|\hat{\theta}_j, s_j; G\right), \tag{87}$$

with the posterior $\mathcal{P}$ defined as in (10). A non-parametric estimate of this quantity derived from log-spline deconvolution is:

$$\hat{\theta}_j^* = \int \theta d\mathcal{P}\left(\theta|\hat{\theta}_j, s_j; \hat{G}_{LS}\right) = \frac{\sum_{m=1}^{M} \bar{\theta}_m \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j - \bar{\theta}_m}{s_j}\right) g_m(\hat{\alpha})}{\sum_{m=1}^{M} \frac{1}{s_j}\phi\left(\frac{\hat{\theta}_j - \bar{\theta}_m}{s_j}\right) g_m(\hat{\alpha})}, \tag{88}$$

where $\hat{\alpha}$ is calculated according to equation (86).

Changing notation relative to Section 2.1, we can contrast the non-parametric shrinkage estimate $\hat{\theta}_j^*$ with a linear shrinkage estimate given by:

$$\hat{\theta}_j^{lin} = \left(\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2 + s_j^2}\right)\hat{\theta}_j + \left(\frac{s_j^2}{\hat{\sigma}_\theta^2 + s_j^2}\right)\hat{\mu}_\theta, \tag{89}$$

where $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta^2$ are estimates of the mean and variance of G. These hyperparameters can be calculated either using simple mean and variance estimators as in Sections 2.1 and 3.1, or based on the first two moments of a non-parametric $\hat{G}$.

The choice between $\hat{\theta}_j^{lin}$ and $\hat{\theta}_j^*$ mirrors the usual tradeoff between the robustness and simplicity of linear estimators versus potential efficiency gains from non-linear estimators. If the mixing distribution is far from a normal distribution the non-parametric approach will leverage the higher moments of G to produce a richer posterior distribution, reducing mean squared error relative to linear shrinkage. On the other hand, non-parametric deconvolution requires estimating extra parameters, and the higher moments of G may be poorly estimated. Linear shrinkage provides a minimum MSE linear approximation to the posterior mean without the need to estimate these higher moments. In applications of non-parametric EB for value-added estimation, comparing non-parametric and linear shrinkage posteriors is a useful reality check on the output of non-linear shrinkage procedures.

Incorporating precision-dependence

The general mixing distribution model in equation (83) assumes independence between effect sizes and standard errors. The same precision-dependence issues discussed in Section 2.6 apply to non-parametric estimation of this model. Like the precision-weighted mean and variance estimators in equations (53) and (54), the maximum likelihood estimators in (84) and (86) will leverage all the restrictions implied by independence of θ_j and s_j to increase efficiency. Imposing these restrictions will enhance precision if they are satisfied but may compromise consistency of the resulting $\hat{G}$ if not. The performance of non-parametric EB posteriors may also suffer if the prior erroneously rules out precision-dependence.

Paralleling the parametric approach of Section 2.6, we can build precision-dependence into non-parametric EB by estimating a model of the relationship between effect sizes and standard errors, then deconvolving residuals. A non-parametric extension of the conditional location/scale model

in equation (55) is given by:

$$\theta_j = \mu(s_j) + \sigma(s_j)r_j,\ r_j|s_j \sim G_r, \tag{90}$$

with $E[r_j|s_j] = 0$ and $Var(r_j|s_j) = 1$. Chen (2023) proposes a CLOSE-NPMLE estimator that estimates the conditional mean and variance functions $\mu(s_j)$ and $\sigma(s_j)$ by non-parametric (local linear) regression, then applies NPMLE to the resulting residuals to estimate G_r. His theoretical results establish that this estimator yields EB decision rules that approximate the decisions of an oracle who knows the full conditional mixing distribution, even if the conditional location/scale model is misspecified. A tractable semi-parametric alternative (sacrificing some flexibility in the first step) is to specify functional forms for $\mu(s_j)$ and $\sigma(s_j)$, estimate the unknown parameters of these functions, then non-parametrically deconvolve residuals with NPMLE or log-spline deconvolution. I provide an example of such an approach in Section 3.7.

3.3 Partial Identification

The normal noise model (82) is usually justified as an asymptotic approximation with a growing number of observations in each group. With few observations per group the distribution of the noise in $\hat{\theta}_j$ will depend on the distribution of the underlying microdata Y_i, and the mixing distribution G may not be non-parametrically identified. However, it may still be possible to recover useful features of the prior and posterior distributions with a partial identification approach. I illustrate such an approach through an example drawn from Kline and Walters' (2021) study of job-specific employment discrimination.

Example 5: Job-level employment discrimination

Kline and Walters (2021) analyze data from resume correspondence experiments sending fictitious applications to real job vacancies (Bertrand and Mullainathan, 2004; Arceo-Gomez and Campos-Vasquez, 2014; Nunley et al., 2015). To manipulate employers' perceptions of race, resumes in these experiments are assigned racially-distinctive names. Suppose each vacancy $j \in \{1, ..., J\}$ in a study receives L applications, with race assignment stratified so that L_w have distinctively-white names and $L_b = L - L_w$ have distinctively-Black names. For example, Bertrand and Mullainathan (2004) sent four applications per job with two in each racial group, so $L_w = L_b = 2$.

Let $D_i \in \{1, ..., J\}$ denote the vacancy that received application i, let $R_i \in \{w, b\}$ represent the racial valence of the name assigned to this application (white or Black), and let $Y_i \in \{0, 1\}$ denote an indicator equal to one if the application was called back by the employer. Assume callback outcomes are generated by Bernoulli trials with stable job-by-race callback probabilities:

$$Y_i|D_i = j, R_i = r \sim Bernoulli(p_{jr}). \tag{91}$$

The unknown unit-specific parameter in this case is the 2×1 vector $\theta_j = (p_{jw}, p_{jb})'$. As in Example 3, this model of independent Bernoulli trials implies the success counts $C_{jr} = \sum_{i=1}^{N} 1\{D_i = j\}1\{R_i = r\}Y_i$ follow $Bin(L_r, p_{jr})$ distributions for $r \in \{w, b\}$.

Adopting an empirical Bayes view, suppose the two job-specific success probabilities are drawn randomly from a bivariate mixing distribution G:

$$(p_{jw}, p_{jb})' \sim G. \tag{92}$$

This mixing distribution describes heterogeneity in discrimination across jobs in the population. We would like to use the distribution of observed success counts to learn about G. However, it is clear that an asymptotic normal approximation does not apply with only two trials per group, so the assumptions underlying the deconvolution methods discussed in Section (3.2) do not hold.

What can we learn about the mixing distribution in this case? With binomial trials, the likelihood of a particular callback configuration (c_w, c_b) conditional on the callback probabilities at job j is:

$$\begin{aligned} f(c_w, c_b|p_{jw}, p_{jb}) &\equiv \Pr[C_{jw} = c_w, C_{jb} = c_b|p_{jw}, p_{jb}] \\ &= \left(\begin{array}{c} L_w \\ c_w \end{array} \right) \left(\begin{array}{c} L_b \\ c_b \end{array} \right) p_{jw}^{c_w}(1-p_{jw})^{L_w - c_w} p_{jb}^{c_b}(1-p_{jb})^{L_b - c_b}. \end{aligned} \tag{93}$$

Combined with the mixing distribution model in (92), equation (93) implies the share of jobs with this callback configuration is given by:

$$\begin{aligned} \bar{f}(c_w, c_b) &\equiv \Pr[C_{jw} = c_w, C_{jb} = c_b] \\ &= \int f(c_w, c_b|p_w, p_b) dG(p_w, p_b) \end{aligned}$$

$$= \left(\begin{array}{c} L_w \\ c_w \end{array} \right) \left(\begin{array}{c} L_b \\ c_b \end{array} \right) \sum_{k=0}^{L_w} \sum_{m=0}^{L_b} (-1)^{k+m} \left(\begin{array}{c} L_w - c_w \\ k \end{array} \right) \left(\begin{array}{c} L_b - c_b \\ m \end{array} \right) \mu(c_w + k, c_b + m), \tag{94}$$

where the function $\mu(x, y) = \int p_w^x p_b^y dG(p_w, p_b)$ describes non-central moments of the bivariate distribution G. Evaluating this expression for all observed values of (c_w, c_b) yields a linear system of the form $\bar{f} = B\mu_G$, which links a vector of callback frequencies $\bar{f}$ to a vector of moments μ_G of the mixing distribution via a known invertible matrix B of binomial coefficients. We can then solve for $\mu_G = B^{-1}\bar{f}$ to recover a set of moments of G from the observed callback frequencies.

This argument shows that all moments of G involving powers of (p_w, p_b) up to (L_w, L_b) are identified. As a result, useful measures of heterogeneity in discrimination are identified even with few trials per job. For example, the variance of the white/Black gap in callback probabilities involves moments up to order two:

$$Var(p_{jw} - p_{jb}) = \mu(2,0) - \mu(1,0)^2 + \mu(0,2) - \mu(0,1)^2 - 2[\mu(1,1) - \mu(1,0)\mu(0,1)]. \tag{95}$$

The moments in equation (95) are identified in an experiment that sends at least two applications per group (i.e. $\min\{L_w, L_b\} \geq 2$), as in Bertrand and Mullainathan (2004). Such an experiment can therefore be used to distinguish a scenario in which a positive average effect of white names comes

from an equal advantage for white applicants at all jobs (in which case $Var(p_{jw} - p_{jb}) = 0$) from a scenario in which some jobs favor white applicants much more than others (in which case $Var(p_{jw} - p_{jb})$ is large). Intuitively, we should see more modest imbalances of $C_{jw} - C_{jb} = 1$ in the former case, while the latter case generates both more highly-imbalanced jobs with $(C_{jw}, C_{jb}) = (2, 0)$ and more instances of equal treatment with $C_{jw} = C_{jb}$. Kline and Walters (2021) apply this approach to document substantial variation in discrimination in several correspondence experiments.

In addition to characterizing heterogeneity in discrimination, the moments of G imply bounds on features of the prior and posterior distributions that are not point-identified. Consider jobs that call back both applicants with distinctively-white names and no applicants with distinctively-Black names in the Bertrand and Mullainathan (2004) study. How sure should we be that jobs with this seemingly-suspicious callback configuration are discriminating? The share of such jobs that have different callback rates for white and Black applicants is:

$$\Pr[p_{jw} \neq p_{jb} | C_{jw} = 2, C_{jb} = 0] = \frac{\int f(2,0|p_w, p_b) 1\{p_w \neq p_b\} dG(p_w, p_b)}{\bar{f}(2,0)}. \tag{96}$$

The smallest value of this posterior probability that is consistent with the observed callback frequencies $\bar{f}$ solves the problem

$$\min_{G \in \mathcal{G}} \frac{\int f(2,0|p_w, p_b) 1\{p_w \neq p_b\} dG(p_w, p_b)}{\bar{f}(2,0)} \text{ s.t. } \bar{f} = B\mu_G, \tag{97}$$

where $\mathcal{G}$ is the set of all bivariate cumulative distribution functions with domain $[0,1]^2$.

Equation (97) defines an optimization problem that is linear in the probability mass function associated with G, which can be solved with linear programming techniques. Kline and Walters (2021) plug empirical estimates of the callback frequencies $\bar{f}$ into this problem and optimize over discrete mixing distributions defined on a fine two-dimensional grid of callback probabilities. Their results show that experiments with few observations per job can generate informative posterior bounds. For instance, in the Bertrand and Mullainathan (2004) experiment, at least 72% of jobs that call back two white applicants and no Black applicants must be discriminating, even using the most conservative mixing distribution consistent with the experimental data. The corresponding bound for a more recent experiment conducted by Nunley et al. (2015) is 85%. The idea of using a maximally-conservative empirical prior to bound posterior probabilities in this way is closely related to empirical Bayes approaches to multiple testing, as I discuss next.

3.4 EB for Multiple Testing: Large-Scale Inference

Non-parametric empirical Bayes methods are tightly connected with multiple testing problems that arise frequently in labor economics and other areas of applied work. Efron (2012) uses "large-scale inference" to refer to EB methods in this context. Consider a list of null hypotheses for each of J units, such as $H_0 : \theta_j = 0$. This collection of hypotheses might concern which subgroups are affected by an intervention, which schools have value-added below some quantile of the distribution (as in Section 2.5), or which jobs discriminate against racially-distinctive names (as in Section 3.3).

Let $T_j = 1\{\theta_j = 0\}$ denote an indicator equal to 1 if the null is true for unit j. Suppose we conduct an independent test of H_0 for each unit to generate a collection of p-values p_j. For a rejection threshold $\bar{p}$, indicators $\delta_j = 1\{p_j \leq \bar{p}\}$ describe which null hypotheses are rejected.

Traditional hypothesis testing seeks to control the probability of type I error (size) for a single test. In other words, we limit the likelihood of a mistaken rejection when the null is true by adopting a decision rule such that $\Pr[\delta_j = 1|T_j = 1] \leq \alpha$ for a tolerance α. With p-values that are uniformly distributed under the null size is controlled by setting a rejection threshold of $\bar{p} = \alpha$. A concern with multiple tests is that applying this rule may lead to scenarios in which many rejected hypotheses are true. In the extreme case where all null hypotheses are true, all rejected hypotheses will be true, but we will be very likely to reject some if $\bar{p}$ is fixed and J is large. This motivates approaches that control alternative notions of aggregate error such as the family-wise error rate (FWER), which is the probability of at least one mistaken rejection: $FWER = \Pr[\sum_{j=1}^{J} T_j \delta_j \geq 1]$. A standard Bonferroni correction uses a modified rejection threshold $\bar{p} = \alpha/J$ which shrinks with J and thereby controls FWER at level α.

Approaches that control FWER (or generalizations such as k-FWER; Lehmann and Romano, 2005) are natural when each mistaken rejection is very costly. In settings with large numbers of tests, however, controlling the absolute number of mistakes is stringent, and it is often more natural to control the *rate* of mistakes. The *false discovery proportion* (FDP) is given by:

$$FDP = \begin{cases} \frac{\sum_{j=1}^{J} \delta_j T_j}{\sum_{j=1}^{J} \delta_j}, & \sum_{j=1}^{J} \delta_j > 0; \\ 0, & \sum_{j=1}^{J} \delta_j = 0. \end{cases} \tag{98}$$

FDP is the share of rejected hypotheses that are true when we reject at least one hypothesis, and is defined to be zero otherwise. Benjamini and Hochberg (1995) propose controlling the *false discovery rate* (FDR), which is the expectation of FDP:

$$FDR = E[FDP]. \tag{99}$$

FDR gives the expected share of true nulls among rejected hypotheses. With a test procedure that controls FDR, therefore, we should expect most rejected null hypotheses to be false.

An EB approach to FDR control

An empirical Bayes approach facilitates FDR control in a setting with many tests. If we adopt the random effects model (6), each true null indicator T_j is a function of a random draw from G. Under this model, the expected share of true nulls among rejected hypotheses (those with $p_j \leq \bar{p}$)

is

$$\begin{aligned} FDR &= \Pr\left[T_j = 1 | \delta_j = 1\right] \\ &= \frac{\Pr[p_j \leq \bar{p} | T_j = 1] \Pr\left[T_j = 1\right]}{\Pr\left[p_j \leq \bar{p}\right]} \\ &= \frac{\bar{p}\pi_0}{F(\bar{p})}, \end{aligned} \tag{100}$$

where the second line uses Bayes' rule, the third uses that p-values are uniformly distributed under the null, $F(p) = \Pr[p_j \leq p]$ is the marginal CDF of p-values, and $\pi_0 = \int 1\{\theta = 0\} dG(\theta)$.

Equation (100) shows that controlling FDR requires choosing $\bar{p}$ to limit $\bar{p}\pi_0/F(\bar{p})$. The CDF $F(\cdot)$ can be estimated using the observed p-value distribution, e.g. with $\hat{F}(\bar{p}) = J^{-1} \sum_j 1\{p_j \leq \bar{p}\}$. The key unknown quantity in equation (100) is π_0, the population share of true nulls. Importantly, π_0 is a feature of G, and therefore reflects an objective fact about the distribution of parameters in the population being studied. If $\pi_0 = 1$ all hypotheses are true, and any rejection is a mistake. When $\pi_0 = 0$ all rejected hypotheses are false, but so are all hypotheses that are not rejected. In between, π_0 provides the correct prior presumption that the null is true in this population.

Benjamini and Hochberg (1995) propose a conservative approach to FDR control that plugs $\pi_0 = 1$ into equation (100). This approach may still allow FDR control with a $\bar{p}$ greater than zero if the p-values are concentrated toward the origin, so that $F(\bar{p}) >> \bar{p}$. It is clear, however, that we can do better. It is logically inconsistent to assume $\pi_0 = 1$ while finding that $F(\bar{p}) > \bar{p}$. More generally, the probability π_0 is a feature of G, and we can learn about G via empirical Bayes deconvolution. This suggests leveraging the distribution of results across tests to discipline the choice of π_0.

The mixing distribution G is non-parametrically identified in the normal noise model (82)-(83), which suggests π_0 may be point identified in some applications. However, it is standard to treat this probability as partially identified in the multiple testing context. Intuitively, it is difficult to empirically distinguish between G's with mass points at exactly zero or only very close to zero. To bound π_0, note that the marginal density of the p-value distribution at a point p can be written:

$$f(p) = \pi_0 + (1 - \pi_0) f_a(p), \tag{101}$$

where $f_a(p)$ is the density of p-values among false null hypotheses (those with $T_j = 0$). Equation (102) reveals that the observed p-value distribution is a mixture of a uniform density with weight π_0 and an alternative density $f_a(p)$ with weight $1 - \pi_0$ (Efron et al., 2001). While the exact form of $f_a(p)$ is usually unknown, this density cannot be negative, so the marginal p-value density at any point p provides an upper bound on π_0 :

$$f(p) = \pi_0 + (1 - \pi_0) f_a(p) \geq \pi_0 \ \forall p. \tag{102}$$

A useful test should generate p-values that are concentrated toward zero when the null is false,

so we expect $f_a(p)$ to be lowest (and the resulting bound on π_0 to be tightest) in the upper tail of the p-value distribution. Storey (2002) proposes to bound π_0 with the estimator

$$\hat{\pi}_0 = \frac{\sum_{j=1}^{J} 1\{p_j > b\}}{(1-b)J}, \tag{103}$$

where the threshold $b \in [0, 1)$ is a tuning parameter. This amounts to assuming all p-values above b correspond to true nulls and using the share of units in this region to estimate the height of the null density. A higher b results in a tighter bound but noisier estimate. Storey et al. (2004) propose a bootstrap approach to selecting b that balances this tradeoff to minimize MSE. A simple alternative is to select a value of b *a priori* since any b provides an upper bound on π_0.

With an estimated upper bound $\hat{\pi}_0$ in hand, equation (100) gives an upper bound on FDR for any rejection cutoff $\bar{p}$. Evaluating this expression at each observed p-value yields a list of *q-values*, given by:

$$q_j = \frac{p_j \hat{\pi}_0}{\hat{F}(p_j)}. \tag{104}$$

The q-value is an empirical Bayes analogue of the p-value (Storey, 2003).[24] Rather than controlling $\Pr[\delta_j = 1 | T_j = 1]$ with the p-value, we borrow strength from the ensemble of tests to flip the conditioning and control $\Pr[T_j = 1 | \delta_j = 1]$ with the q-value. If we reject all hypotheses with p-values less than p_j, we should expect at most a share q_j of rejections to be mistakes.

3.5 Ranking Problems

Empirical Bayes and related methods are increasingly used to build *report cards* that summarize estimates of quality with rankings or coarse grades. Examples include evalutions of K-12 schools, teachers, colleges, hospitals, doctors, and neighborhoods (Angrist et al., 2024b; Bergman and Hill, 2018; Chetty et al., 2017; Kolstad, 2013; Pope, 2009; Chetty and Hendren, 2018). Report cards of this sort lead naturally to comparisons of top and bottom performers (the "league table mentality" discussed by Gu and Koenker, 2023b). However, the estimators and decision rules I have considered so far aim for good average performance across units, not for accurate pairwise comparisons. Can we be sure that units assigned top grades in a value-added report card are actually among the best?

One approach to this question is to analyze relative rankings of units in a standard frequentist inference framework. Treating the θ_j's as unknown fixed parameters, the rank of unit j is given by $rank_j = \sum_{k=1}^{J} 1\{\theta_j \leq \theta_k\}$. Mogstad et al. (2023) develop methods to construct confidence sets for any individual $rank_j$ as well as simultaneous confidence sets for the rankings of all J units. Andrews et al. (2023) propose tools for inference on the value-added of the highest-ranked unit, accounting for upward bias in the maximum estimate due to the ranking step.

An alternative strategy is to formalize the objective of a report card system in an EB compound decision framework. In some cases this makes clear that mis-ranking units may not be a problem. If our goal is to communicate reliable information on absolute quality, ranking units with an absolute

[24] The approach to q-value estimation outlined here is implemented in the **qvalue** R package (Storey, 2015).

performance metric such as an EB posterior mean may lead to good decisions, even if such a ranking yields a high rate of mistakes in relative comparisons (i.e. a low correlation between reported rankings and the true $rank_j$'s). Such mistakes will be hardest to avoid when the θ_j's are very close together, in which case mis-rankings may be of little consequence for outcomes. In other cases we may be specifically interested in getting the rankings right, which requires a loss function tailored to this goal.

Kline et al. (2024) use such a loss function to construct a report card summarizing firm-specific discrimination estimates.[25] Consider a decision-maker tasked with assigning a grade $\delta_j \in \{1, ..., J\}$ to each of J units. Let $\delta = (\delta_1, ..., \delta_J)'$ denote the vector of grades for all units, and let $\Theta = (\theta_1, ..., \theta_J)'$ denote the vector of true parameters. Suppose the decisionmaker seeks to minimize the loss function

$$\begin{aligned}\mathcal{L}(\delta; \Theta) &= \binom{J}{2}^{-1} \sum_{j=2}^{J} \sum_{k=1}^{j} [1\{\theta_j > \theta_k\} 1\{\delta_j < \delta_k\} + 1\{\theta_j < \theta_k\} 1\{\delta_j > \delta_k\} \\ &\quad -\lambda(1\{\theta_j > \theta_k\} 1\{\delta_j > \delta_k\} + 1\{\theta_j < \theta_k\} 1\{\delta_j < \delta_k\})]. \end{aligned} \tag{105}$$

This function assigns a loss of 1 for each pair of units that are mis-ordered (*discordances*), and a gain of $\lambda \in [0, 1]$ for each pair that is correctly ordered (*concordances*). Assigning two units the same grade ($\delta_j = \delta_k$) guarantees a loss of zero for the pair.

By rearranging terms in equation (105), we can represent this loss function as

$$\mathcal{L}(\delta; \Theta) = (1 - \lambda) DP(\delta; \Theta) - \lambda \tau(\delta; \Theta), \tag{106}$$

where $DP(\delta; \Theta) = \binom{J}{2}^{-1} \sum_{j=2}^{J} \sum_{k=1}^{j} [1\{\theta_j > \theta_k\} 1\{\delta_j < \delta_k\} + 1\{\theta_j < \theta_k\} 1\{\delta_j > \delta_k\}]$ is the *discordance proportion* – the share of pairwise comparisons that are incorrect – and $\tau(\delta; \Theta)$ is Kendall's τ measure of rank correlation between grades δ_j and parameters θ_j (the share of concordances minus the share of discordances). This expression shows that a decision-maker who weights concordances and discordances equally ($\lambda = 1$) will seek to maximize the rank correlation between true parameters and grades. A λ below one can be interpreted as *discordance aversion* that penalizes mistakes over and above their effect on the rank correlation. Such discordance aversion creates an incentive to coarsen rankings and create a report card with fewer than J grades. If ranking mistakes are very costly, it may be optimal to group units together and avoid making assertions about relative performance of units that cannot be distinguished with sufficient certainty.

Following the approach of Section 2.5, an empirical Bayes grading system treats the θ_j's as random draws from a mixing distribution G, and minimizes risk based on a deconvolution estimate $\hat{G}$. Optimal decisions for an oracle that knows G are given by

$$\delta^* = \arg\min_{\delta} \ (1 - \lambda) DR(\delta) - \lambda \bar{\tau}(\delta), \tag{107}$$

[25] See Gu and Koenker (2022) and Gu and Koenker (2023b) for other recent approaches to ranking and selection in an EB framework.

where the *discordance rate* $DR(\delta)$ is the posterior expectation of the discordance proportion $DP(\delta;\Theta)$, and $\bar{\tau}(\delta)$ is the posterior expectation of the rank correlation $\tau(\delta;\Theta)$. Under model (82)-(83) with a continuous G, these quantities can be written

$$\bar{\tau}(\delta) = \left(\begin{array}{c} J \\ 2 \end{array} \right) \sum_{j=2}^{J} \sum_{k=1}^{j} \left[2\pi_{jk} - 1\right] \left[1\{\delta_j > \delta_k\} - 1\{\delta_j < \delta_k\}\right], \tag{108}$$

$$DR(\delta) = \left(\begin{array}{c} J \\ 2 \end{array} \right) \sum_{j=2}^{J} \sum_{k=1}^{j} \left[\pi_{jk} 1\{\delta_j < \delta_k\} + (1 - \pi_{jk}) 1\{\delta_j > \delta_k\}\right], \tag{109}$$

where $\pi_{jk} = \Pr[\theta_j > \theta_k | \hat{\theta}_j, \hat{\theta}_k, s_j, s_k]$ is the posterior probability that value-added for unit j exceeds that of unit k given the estimates and standard errors for both units:

$$\pi_{jk} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{t} \frac{1}{s_j} \phi\left(\frac{\hat{\theta}_j - t}{s_j}\right) \frac{1}{s_j} \phi\left(\frac{\hat{\theta}_k - u}{s_k}\right) dG(u) dG(t)}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{s_j} \phi\left(\frac{\hat{\theta}_j - t}{s_j}\right) \frac{1}{s_j} \phi\left(\frac{\hat{\theta}_k - u}{s_k}\right) dG(u) dG(t)}. \tag{110}$$

Equations (108) and (109) show that the decision-relevant features of the posterior distribution for this ranking problem are the pairwise posterior probabilities π_{jk}. With two units the optimal grading rule sets $\delta_j > \delta_k$ if $\pi_{jk} > (1+\lambda)^{-1}$, sets $\delta_j < \delta_k$ if $\pi_{jk} < \lambda(1+\lambda)^{-1}$, and declares a tie ($\delta_j = \delta_k$) otherwise. With more than two units such pairwise decisions may lead to Condorcet (1785)-style cycles that violate transitivity – for example, for three units j, k, and m, the decision-maker may want to set $\delta_j > \delta_k$ but $\delta_j = \delta_m$ and $\delta_k = \delta_m$. Kline et al. (2024) represent problem (107) as an integer linear programming problem with transitivity constraints that rule out such cycles. An EB solution $\hat{\delta}^*$ plugs empirical posterior probabilities $\hat{\pi}_{jk}$ into this problem based on a mixing distribution estimate $\hat{G}$ from the deconvolution step. The resulting report card classifies units into coarse grades that balance information content (captured by $\bar{\tau}(\hat{\delta}^*)$) against ranking mistakes (captured by $DR(\hat{\delta}^*)$), with the strength of this tradeoff determined by the preference parameter λ.

3.6 Compound Decisions and Shrinkage Strategies

I next synthesize themes from throughout the chapter in a simple compound decision framework.[26] Consider a decision-maker who aims to select units with high values of a parameter θ_j, which is assumed to be positive for all units ($\theta_j \geq 0$). For example, θ_j might be value-added of unit j relative to a baseline or status-quo reference category. The decision-maker earns utility $\theta_j^{1-\omega}$ if she selects a unit with parameter θ_j and pays a constant cost κ for each selection. Parameter $\omega \in [0, 1)$ governs the decision-maker's risk aversion, with $\omega = 0$ corresponding to risk-neutrality. The component-wise loss function for this decision-maker is given by:

$$\ell(\theta_j, \delta_j) = -\delta_j \left(\theta_j^{1-\omega} - \kappa\right), \tag{111}$$

[26] See Section XI.A of Kline et al. (2022) for related discussion in the context of auditing to detect discrimination.

where $\delta_j \in \{0, 1\}$ indicates selection of unit j.

Suppose the decision-maker has access to estimates and standard errors $(\hat{\theta}_j, s_j)$ for each unit to use in making decisions. These estimates are unbiased and normally distributed as in equation (82). The underlying parameters θ_j are drawn from a distribution G according to model (83). With J units, the risk of a decision rule δ is then

$$\mathcal{R}(\delta; G, \omega) = -\sum_{j=1}^{J} \int \int \delta(\hat{\theta}, s_j) \left(\theta_j^{1-\omega} - \kappa\right) \frac{1}{s_j} \phi\left(\frac{\hat{\theta} - \theta}{s_j}\right) d\hat{\theta} dG(\theta). \tag{112}$$

I have written $\mathcal{R}(\delta; G, \omega)$ as a function of G and ω to emphasize that risk depends on the mixing distribution as well as the decision-maker's preferences. The optimal decision rule for an oracle that knows G is $\delta^*(G, \omega) = \arg\min_{\delta \in \mathcal{D}} \mathcal{R}(\delta; G, \omega)$, where $\mathcal{D}$ is the set of candidate decision rules mapping $(\hat{\theta}_j, s_j)$ to binary selection decisions. It is straightforward to show that such an oracle selects units if and only if the posterior expectation of $\theta_j^{1-\omega}$ exceeds the cost κ:

$$\delta^*(\hat{\theta}_j, s_j; G, \omega) = 1\left\{\int \theta^{1-\omega} d\mathcal{P}(\theta|\hat{\theta}_j, s_j; G) \geq \kappa\right\}, \tag{113}$$

where the posterior distribution $\mathcal{P}$ is defined as in equation (10).

Equation (113) facilitates comparison of shrinkage approaches distinguished by different degrees of risk aversion. When $\omega = 0$ the decision-maker selects units based on a cutoff in the posterior mean θ_j^*. As ω grows, the expectation that determines the decision rule becomes increasingly sensitive to the lower tail of the posterior distribution. In the limit as ω approaches 1, we have

$$\lim_{\omega \to 1} \delta^*(\hat{\theta}_j, s_j; G, \omega) = 1\left\{\Pr[\theta_j = 0|\hat{\theta}_j, s_j] \leq 1 - \kappa\right\}. \tag{114}$$

The probability $\Pr[\theta_j = 0|\hat{\theta}_j, s_j] = \int 1\{\theta = 0\} d\mathcal{P}(\theta|\hat{\theta}_j, s_j; G)$ is the share of units with $\theta_j = 0$ among those with a particular estimate and standard error, also known as a *local false discovery rate* (Efron et al., 2001). Equation (114) shows that a maximally risk-averse decision-maker selects units based on a cutoff in LFDR. We can therefore think of decisions based on posterior means and false discovery rates as endpoints of a continuum of shrinkage strategies traced out by varying risk aversion.

A decision-maker who does not know G must estimate the posterior expectation in equation (113) to implement a feasible EB decision rule. When G is point-identified as in model (82)-(83), it is sensible to estimate the mixing distribution with the deconvolution methods described in Section 3.2. When the mixing distribution is partially identified as in Section 3.3, there are multiple G's that are consistent with the distribution of the observed data, and it is no longer clear which one to use for decision-making. *Minimax* decisions provide an important conservative benchmark in this case (Wald, 1945; Savage, 1951; Manski, 2000). Formally, a minimax decision rule is given by:

$$\delta^{mm}(\omega) = \arg\min_{\delta \in \mathcal{D}} \max_{G \in \mathcal{G}_\mathcal{I}} \mathcal{R}(\delta; G, \omega), \tag{115}$$

where $\mathcal{G}_\mathcal{I}$ is the identified set for the mixing distribution (i.e. the set of G's consistent with the population distribution of the observed data). Problem (115) can be motivated by an adversarial setup in which an opponent observes the decision rule selected by the analyst and chooses the maximally-damaging mixing distribution in response. An EB minimax rule plugs a deconvolution estimate of the identified set $\mathcal{G}_\mathcal{I}$ into (115).

This decision framework clarifies the relationship between the EB posterior mean emphasized for much of this chapter and other shrinkage strategies such as the multiple testing approach developed in Section 3.4. The conventional approach of listing units ordered by EB posterior means mimics the selection decisions of a risk-neutral decision-maker who forms posteriors based on a point estimate of G. In contrast, ordering units by q-values is conservative in two senses. First, this approach corresponds to the decisions of a maximally risk-averse decision-maker, as shown in equation (114). Second, by using an estimated upper bound on the prior probability π_0, the q-value approach adopts a worst-case empirical prior that is as favorable as possible to the null hypothesis among those consistent with the data, in the spirit of the minimax rule in (115). Whether one of these two shrinkage strategies (or something in between) is preferable depends on the context and economic goal – is the aim to select units that will improve average outcomes, or to recommend units that are likely to generate improvements even in a worst-case scenario? Section 3.7 presents an empirical contrast of these two types of shrinkage.

3.7 Non-Parametric EB Application: Firm-Level Labor Market Discrimination

This subsection applies non-parametric empirical Bayes methods to study variation in race and gender discrimination across large US employers. My analysis revists a large resume correspondence experiment conducted by Kline, Rose and Walters (2022; 2024), which extended the analysis of Kline and Walters (2021) to study variation in discrimination across entire firms rather than individual jobs. This experiment submitted applications to multiple entry-level job vacancies nested within 108 Fortune 500 employers. Up to 125 vacancies were sampled for each firm, with each vacancy for a given firm in a different US county. Following Bertrand and Mullainathan (2004), resumes were randomly assigned distinctive names to convey race and gender to the employer, with race assignment stratified so that each vacancy received 4 distinctively-Black and 4 distinctively-white names. Male and female names were each assigned to 50% of applications with no stratification. The primary outcome is an indicator for whether the employer attempted to contact an applicant within 30 days by phone or e-mail (I sometimes use "callbacks" as shorthand for these contacts). Following Kline et al. (2024), I focus on 97 firms with at least 40 sampled vacancies and overall callback rates above 3 percent. This results in a sample of 78,910 applications to 10,453 jobs nested within 97 firms. Full details on the experimental design and sample characteristics are available in Kline et al. (2022) and Kline et al. (2024).

My analysis mostly follows Kline et al. (2022), with some departures to illustrate issues discussed in the preceding sections. In step 1 of the EB recipe, I estimate firm-level discrimination parameters and corresponding standard errors for each employer. These estimates come from firm-specific OLS

regressions of callbacks on race, which can be written:

$$Y_i = \sum_{j=1}^{J} D_{ij}[\alpha_j + \theta_j 1\{R_i = w\}] + e_i, \tag{116}$$

where Y_i indicates a callback for application i, D_{ij} indicates that application i was sent to a vacancy at firm j, and $R_i \in \{w, b\}$ denotes the racial distinctiveness of the name assigned to the application. Parameter α_j measures the callback rate for applications with distinctively-Black names at firm j, while θ_j captures the gap in contact rates between distinctively-white and distinctively-Black names at this firm. Since race was randomly assigned we can interpret θ_j as an average treatment effect of distinctively-white names on callbacks at firm j.[27] Corresponding OLS estimates $\hat{\theta}_j$ provide unbiased estimates of these causal parameters. Standard errors s_j are clustered by job to account for job-level differences in overall contact rates along with stratification of race assignments by job. Models for gender replace the race indicator with an indicator for a distinctively-male name in equation (116).

Distributions of discrimination

Step 2 of the EB recipe uses the firm-specific estimates to summarize the distribution of discrimination across firms. I first report means and bias-corrected standard deviations of race and gender contact gaps. The estimated mean $\hat{\mu}_\theta$ is the average of $\hat{\theta}_j$'s as in equation (8), while the estimated standard deviation is the square root of the bias-corrected variance estimate from equation (80). Since estimates are independent across firms the mixing variance estimator simplifies to $\hat{\sigma}^2_\theta = (J-1)^{-1}\sum_j[(\hat{\theta}_j - \hat{\mu}_\theta)^2 - J^{-1}(J-1)s_j^2]$ in this case. As in Kline et al. (2022), I report standard errors for $\hat{\mu}_\theta$ and $\hat{\sigma}_\theta$ based on a job-clustered weighted bootstrap procedure that draws *iid* exponential weights for each job and reweights the regressions used to produce $(\hat{\theta}_j, s_j)$ in each bootstrap iteration.

Firms favor distinctively-white names over distinctively-Black names on average, and this gap is highly variable across firms. This can be seen in column (1) of Table 1, which shows estimated means and standard deviations of firm-specific OLS race coefficients. The mean level gap in column (1) demonstrates that on average, firms call applications with distinctively-white names 2.1 percentage points more often than applications with distinctively-Black names, an estimate that is highly statistically significant (t-statistic > 11). Subsequent rows compare unadjusted and bias-corrected standard deviations of contact gaps. Bias correction reduces the standard deviation of gaps from 2.4 percentage points to 1.7 percentage points, which implies that $(1-(0.017/0.024)^2)\times 100 = 50\%$ of the variance in $\hat{\theta}_j$'s is due to statistical noise rather than true firm heterogeneity. Nonetheless, the bias-corrected estimate reveals large differences in discrimination even after accounting for this noise: a firm that is one standard deviation above the mean penalizes distinctively-Black names 80 percent more than the average firm.

In contrast to the results for race, column (3) of Table 1 shows that the mean gender coefficient is a precisely-estimated zero. However, the bias-corrected standard deviation estimate is even larger

[27] See the Appendix to Kline et al. (2022) for a potential outcomes framework formalizing this interpretation.

for gender than for race (0.031 versus 0.017). The combination of a zero mean and a large standard deviation implies that there must be mass both above and below zero – some firms favor men, while others favor women. This finding highlights the value of an EB analysis of firm heterogeneity: a focus on the mean would suggest little gender discrimination in this experiment, but the second moment of the distribution reveals substantial discrimination operating in each direction.

I next characterize the full mixing distribution of firm-specific discrimination with the log-spline deconvolution estimator of Efron (2016). Unlike the mean and variance estimates in Table 1, this deconvolution procedure requires taking a stand on the relationship between effect sizes and precision. The basic tests discussed in Section 2.6 reject independence in this case, which indicates that accounting for precision-dependence is likely to be important. Specifically, a regression of the race level gap $\hat{\theta}_j$ on $\log s_j$ yields a coefficient of 0.034 with a robust standard error of 0.005. This indicates strong precision-dependence in the conditional mean of the race gap. A corresponding regression of the male/female gap on its log standard error yields an insignificant coefficient of -0.005 (SE = 0.018), but a regression of $(\hat{\theta}_j - \hat{\mu}_\theta)^2 - s_j^2$ on $\log s_j$ yields a marginally significant coefficient of 0.0035 (SE = 0.0019), suggesting potential dependence in the conditional variance.

Motivated by these tests, I estimate models of dependence between effect sizes and standard errors and deconvolve residuals from these models. Seventy-eight of the 97 estimated racial gaps favor distinctively-white names, and Kline et al. (2022) test and cannot reject the hypothesis that the few observed negative estimates are attributable to sampling error. Following Kline et al. (2024), I therefore adopt a model of precision-dependence that implies white/Black contact gaps are positive for each value of s_j:

$$\theta_j = \exp\left(\psi_1 + \psi_2 \log s_j\right) r_j, \;\; r_j | s_j \sim G_r, \tag{117}$$

where r_j has positive support and $E[r_j] = 1$. This model implies $E[\hat{\theta}_j | s_j] = \exp(\psi_1 + \psi_2 \log s_j)$. I estimate ψ_1 and ψ_2 in a first-step non-linear least squares regression, which yields estimates of $\hat{\psi}_1 = 2.52$ (SE = 0.80) and $\hat{\psi}_2 = 1.56$ (SE = 0.21). I then form residuals $\hat{r}_j = \hat{\theta}_j / \exp(\hat{\psi}_1 + \hat{\psi}_2 \log s_j)$, and estimate G_r by applying the log-spline deconvolution estimator to these residuals, assuming $\hat{r}_j | r_j, s_j \sim \mathcal{N}\left(r_j, \exp(-2\hat{\psi}_1) s_j^{2(1-\hat{\psi}_2)}\right)$ and constraining the mean of the deconvolved distribution to equal 1. The log-spline penalty in equation (86) is calibrated to match a bias-corrected estimate of the variance of r_j. I choose the other log-spline tuning parameters by fixing the number of spline knots at $K = 5$ and using $M = 1,000$ equally-spaced support points between zero and the empirical maximum of $\hat{r}_j$.

Model (117) is inappropriate for gender since we know from Table 1 that some gender gaps are positive while others are negative. The tests discussed above also indicate dependence in the conditional variance but not the conditional mean. I accommodate these findings with the alternative model

$$\theta_j = \psi_0 + s_j^{\psi_2} r_j, \;\; r_j | s_j \sim G_r, \tag{118}$$

where $E[r_j] = 0$ and $Var(r_j) = \sigma_r^2$. This model implies $E[\hat{\theta}_j | s_j] = \psi_0$ and $E[(\hat{\theta}_j - \psi_0)^2 - s_j^2 | s_j] =$

$\sigma_r^2 s_j^{2\psi_2}$. I estimate ψ_0 with the mean of the $\hat{\theta}_j$'s, then estimate ψ_2 by non-linear least squares based on the second moment condition. The resulting estimates are $\hat{\psi}_1 = -0.001$ (SE = 0.005) and $\hat{\psi}_2 = 0.854$ (SE = 0.361). I then deconvolve the residuals $\hat{r}_j = (\hat{\theta}_j - \hat{\psi}_1)/s_j^{\hat{\psi}_2}$ assuming $\hat{r}_j | r_j, s_j \sim \mathcal{N}\left(r_j, s_j^{2(1-\hat{\psi}_2)}\right)$ and constraining the mean of the deconvolved distribution to equal zero. Following the analysis for race, I use a five-knot spline with $1,000$ equally-spaced support points between the empirical minimum and maximum of the estimated residuals.

Results of these flexible deconvolutions appear in Figure 2, with log-spline estimates for race in panel A and estimates for gender in panel B. Each panel displays the deconvolved distribution of the residual r_j along with the resulting marginal distribution of θ_j, which comes from applying a change-of-variables to the distribution of residuals combined with the empirical distribution of standard errors.[28] The first two moments of these distributions are close to the means and variances reported in Table 1, but the non-parametric deconvolutions also reveal more subtle features of the distributions of race and gender discrimination. The distribution of race gaps in panel A(ii) is asymmetric with a long right tail, suggesting that while all firms weakly favor distinctively-white names, the average race gap is driven by a small share of heavy discriminators. The gender gap distribution in panel B(ii) is symmetric with a sharp peak at zero and fat tails in both directions. This indicates that most firms display little preference for male or female names, but a few discriminate substantially against each group.

Posterior predictions of discrimination

Moving to step 3 of the EB recipe, I next consider the prospects for learning about firm-specific discrimination parameters via empirical Bayes shrinkage. Figure 3 plots posterior mean race and gender gaps $\hat{\theta}_j^*$ derived from the log-spline prior estimates, overlaid on the prior distribution and a histogram of the unbiased $\hat{\theta}_j$ estimates. Posteriors for each θ_j are constructed by computing EB posterior mean residuals $\hat{r}_j^*$, then transforming these to predict θ_j according to equations (117) and (118). As usual, the shrunk posteriors are less variable than the mixing distribution, which is less variable than the noisy unbiased estimates. Still, the posterior mean estimates are dispersed enough to generate informative estimates of firm-specific parameters. The standard deviation of the posterior mean estimates for race is 0.014 and the average squared standard error of the $\hat{\theta}_j$'s is 0.0003. Coupled with the mixing standard deviation of 0.018 in Figure 2, this implies shrinkage reduces MSE by an estimated $(1 - [(0.018^2 - 0.014^2)/0.0003]) \times 100 = 57\%$. For gender, the corresponding reduction in MSE is 75%.

Figure 4 assesses differences between non-parametric and linear shrinkage by plotting posterior means $\hat{\theta}_j^*$ against linear shrinkage estimates $\hat{\theta}_j^{lin}$. The linear shrinkage estimates are constructed as in equation (14) using mean and variance hyperparameter estimates from Table 1. The differences between linear and non-linear shrinkage estimates are most evident in the tails of the distribution.

[28] Estimates of the precision-dependence parameters and residual distribution from equations (117) and (118) imply a conditional distribution of θ_j for each observed value of s_j. I construct a smoothed estimate of the marginal distribution of θ_j by creating an equally-spaced grid of M support points between the minimum and maximum support limits of the conditional distributions, and assigning the mass at each point of the conditional distributions to the closest support point in the marginal grid (as measured by the absolute value of the difference between support points).

As shown in panel A(i), the non-parametric posterior means for race inherit the positive support restriction I imposed on the prior, so the few negative point estimates are shrunk to slightly above zero. The non-parametric and linear shrinkage posteriors also differ noticeably in the upper tail, where the non-parametric procedure generates larger values. This is a consequence of two forces: the upper tail of the non-parametric prior distribution is thicker than that of a normal distribution, and the non-parametric posterior incorporates precision-dependence with effect sizes increasing in standard errors. As a result, large positive point estimates are shrunk less by the non-parametric procedure, particularly those with large standard errors.

I parse these explanations in panel A(ii) by incorporating precision-dependence into the linear shrinkage approach. Specifically, I apply linear rather than non-parametric shrinkage to the estimated residuals $\hat{r}_j$ from equation (117) before transforming the residuals to compute posteriors for θ_j. Posteriors from this conditional shrinkage strategy align better with the non-parametric posterior estimates throughout the distribution. However, non-parametric shrinkage still generates some larger posterior means at the top of the distribution due to the thick tail of the non-parametric prior. Panel B shows that non-parametric shrinkage also generates more extreme posterior mean gender gaps in the tails of the distribution than linear shrinkage, especially for one firm with a large negative posterior mean indicating substantial discrimination against men. This is a consequence of the extra mass in the left tail of the prior distribution displayed in Figure 3B(ii).

Multiple testing to detect discrimination

In addition to forming posterior mean estimates with low mean squared error, it is natural ask what we can say with confidence about which firms discriminate against distinctively-Black names at all. I investigate this question with a multiple testing analysis along the lines of Section 3.4. This analysis begins with one-tailed z-tests of the null hypothesis $H_0 : \theta_j = 0$ against the alternative $H_A : \theta_j > 0$, generating p-values $p_j = 1 - \Phi(\hat{\theta}_j / s_j)$ for each firm j.[29]

These tests generate small p-values for many firms. Panel A of Figure 5 displays a histogram of the p-values p_j, which are concentrated toward zero. This is unsurprising since the moment estimates in Table 1 show the mean and variance of G are clearly not zero in this case – this implies that some firms must be discriminating. To obtain a conservative estimate of the share of firms that are *not* discriminating, I set the threshold b in equation (103) to 0.5 (shown as a black vertical line in Figure 5A), which generates an estimated bound of $\hat{\pi}_0 = 0.39$ (the red horizontal line). In other words, at least 61% of firms must be discriminating to rationalize the observed p-value distribution.

This bound on the prior share of true nulls implies that many firms can be reliably classified as discriminating even while controlling the false discovery rate to a low level. I compute q-values by plugging $\hat{\pi}_0$ into equation (104) along with a CDF estimate $\hat{F}(p)$ based on the empirical distribution function of p-values. A histogram of the resulting q-value distribution appears in panel B of Figure 5. Twenty-eight firms have q-values below 0.05. Since a q-value threshold of 0.05 controls FDR

[29] Kline et al. (2022) conduct a similar multiple-testing analysis for the full sample of 108 firms based on paired-sample t-tests. I use the subsample of 97 firms with large samples studied by Kline et al. (2024) and employ a simple z-test for illustrative purposes, which tends to yield slightly smaller p-values.

to 5%, we should expect at most one out of every twenty firms with $q_j \leq 0.05$ to have $\theta_j = 0$ in repeated applications of this test procedure.

Contrasting shrinkage approaches

Table 2 summarizes the empirical Bayes analysis of this experiment by listing the 97 firms in the sample ordered by their q-values. The table reports each firm's unbiased estimate $\hat{\theta}_j$, standard error s_j, p-value p_j, q-value q_j, non-parametric posterior mean $\hat{\theta}_j^*$, and linear shrinkage posterior mean $\hat{\theta}_j^{lin}$. The various EB shrinkage estimates are broadly aligned across firms. For example, Genuine Parts (Napa Auto) has the smallest q-value along with the largest point estimate, the largest linear shrinkage estimate, and the largest non-parametric posterior mean. However, some notable differences are evident across these measures as well. For instance, Walmart has both a large point estimate (0.070) and a large standard error (0.039). Since the estimated prior model indicates strong positive dependence between effects and standard errors, this company's estimate is shrunk very little, and it receives the second-highest posterior mean on the list (0.070). Given its large standard error, however, Walmart receives a q-value of 0.05, ranking 29th by this metric. Conversely, VFC has a modest point estimate of 0.038 but a small standard error of 0.014, which results in both a low q-value of 0.018 and a low posterior mean of 0.022.

Figure 6 presents a more systematic investigation of differences between shrinkage approaches by plotting decision frontiers based on posterior means and q-values. Each decision rule selects the 20% of firms with most extreme estimates of discrimination against distinctively-Black names, corresponding to posterior means above 0.032 and q-values below 0.024. These decision rules can be seen as versions of the selection rule in equation (113), setting $\omega = 0$ for the posterior mean, $\omega \to 1$ for the q-value, and calibrating the cost κ so that 20% of firms are selected. Observed combinations of point estimates and log standard errors are denoted with black points, and shaded regions depict hypothetical decisions for combinations not observed in the experiment.

While there is substantial overlap between firms selected based on posterior means and q-values, there is also slippage between these two decision rules. Specifically, 13 of the 19 firms selected by the posterior mean rule are also selected by the q-value rule. Similarly, since each rule selects the same total number by design, 13 of 19 selected by the q-value rule are also selected by the posterior mean rule. Columns (7) and (8) of Table 2 label the firms selected by each decision rule.

Discrepancies between these classifications arise because of the differing shapes of the selection frontiers depicted in Figure 6. A q-value decision rule defines an upward-sloping frontier in the plane relating point estimates to log standard errors. Since there is more uncertainty for firms with large standard errors, higher point estimates are necessary to classify such firms as discriminating while limiting the false discovery rate. In contrast, the posterior mean selection frontier defines a *downward*-sloping relationship between point estimates and log standard errors.[30] This pattern is

[30]The posterior mean selection rule becomes extremely non-linear for log standard errors below -4.5, suggesting that an enormous point estimate is required to warrant selection in this region. This phenomenon is due to the upper bound on the support of residuals imposed in Panel A(i) of Figure 2. The maximum of the empirical residuals $\hat{r}_j$ used to set this support is 3.4, which implies θ_j can be no bigger than $\exp(\hat{\psi}_1 + \hat{\psi}_2 \log s_j) \times 3.4$. With $\hat{\psi}_1 = 2.52$, $\hat{\psi}_2 = 1.56$, and a decision cutoff of $\hat{\theta}_j^* \geq 0.032$, a firm with $\log s_j \leq -4.61$ cannot be selected regardless of its point estimate. This issue has little impact on empirical selection decisions since no firms have estimates near the highly

driven by the strong positive relationship between effect sizes and standard errors evident in the scatter plot of black points, which is built into the prior distribution used to construct posterior means. As a result, among firms with intermediate point estimates, the q-value rule selects firms with smaller standard errors (the yellow region) while the posterior mean rule selects firms with larger standard errors (the green region).

While the q-value and posterior mean decision rules select different sets of firms, differences in expected outcomes between these approaches turn out to be modest. Average posterior means among firms selected by the posterior mean and q-value rules are 0.043 and 0.037, implying that a q-value cutoff selects firms with only slightly smaller expected discrimination values. Likewise, the highest q-value among firms selected by the posterior mean rule is only 0.071, suggesting that a cutoff in $\hat{\theta}_j^*$ selects firms where the posterior probability of discriminating is also high. These findings indicate that a risk neutral analyst would pay little price for using the decision rule of a risk-averse decision-maker (and vice versa). More generally, these sorts of contrasts between decision rules can help to trace out the frontier of outcomes available with different strinkage strategies and assess the robustness of EB shrinkage analyses to the assumed form of decision-maker preferences.

4 Conclusion

Empirical research in labor economics increasingly focuses on variation in quality or conduct across large sets of units like firms, schools, or neighborhoods. This chapter has reviewed empirical Bayes methods for quantifying heterogeneity, estimating unit-specific parameters, and making statistical decisions in such studies. The EB recipe outlined here proceeds by estimating each unit's parameter, using the ensemble of parameter estimates to construct an empirical prior distribution, and forming posteriors based on this prior combined with the unit-specific estimates. This EB shrinkage approach can be applied in service of a variety of statistical and economic goals, including reducing aggregate mean squared error, directing workers or consumers to units with favorable expected outcomes, ranking units, and making selection decisions while limiting the likelihood of mistakes.

Thanks to the increasing availability of large-scale administrative labor market data, potential applications of EB methods in labor economics should continue to grow. This is likely to generate avenues for answering novel economic questions as well as new methodological challenges. Most EB value-added analyses in labor economics to date have employed simple James/Stein-style linear shrinkage strategies. However, realistic empirical applications often bear little resemblance to the stylized James/Stein framework with normally-distributed estimates, homogeneous variances, and independence across units. The non-parametric EB methods discussed in the second half of this chapter provide tools for quantifying heterogeneity and implementing shrinkage in more general settings. Applying these methods to account for the complexities of real-world labor market data and research designs is a fruitful direction for empirical work.

A second promising direction is to tighten the link between economic objectives and econometric

non-linear portion of the decision frontier.

estimation when applying EB methods. This chapter has emphasized the connection between EB shrinkage and decision problems that aim to minimize various forms of aggregate error across units. Conventional linear shrinkage is useful for reducing mean squared error and making risk-neutral selection decisions, which may not correspond to how EB estimates will be used in practice. An explicit statement of the relevant loss function clarifies the appropriate shrinkage strategy and injects economic reasoning into value-added analysis. Such an approach has the potential to make EB applications in labor economics more useful and actionable for policymakers, workers, and households.

References

Abadie, A. and M. Kasy (2019): "Choosing among regularized estimators in empirical economics: the risk of machine learning," *The Review of Economics and Statistics*, 101, 743–762.

Abaluck, J., M. Caceres Bravo, P. Hull, and A. Starc (2021): "Mortality effects and choice across private health insurance plans," *The Quarterly Journal of Economics*, 136, 1557–1610.

Abdulkadiroğlu, A., P. A. Pathak, J. Schellenberg, and C. R. Walters (2020): "Do parents value school effectiveness?" *American Economic Review*, 110, 1502–39.

Abdulkadiroğlu, A., J. D. Angrist, S. Dynarski, T. J. Kane, and P. A. Pathak (2011): "Accountability and flexibility in public schools: evidence from Boston's charters and pilots," *Quarterly Journal of Economics*, 126(2), 699–748.

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999): "High wage workers and high wage firms," *Econometrica*, 67, 251–333.

Andrews, I., T. Kitagawa, and A. McCloskey (2023): "Inference on winners," *The Quarterly Journal of Economics*, 139, 305–358.

Andrews, M. J., L. Gill, T. Schank, and R. Upward (2008): "High wage workers and low wage firms: negative assortative matching or limited mobility bias?" *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 673–697.

Angrist, J., P. Hull, P. A. Pathak, and C. Walters (2024a): "Credible school value-added with undersubscribed school lotteries," *The Review of Economics and Statistics*, 106, 1–19.

Angrist, J., P. Hull, P. A. Pathak, and C. R. Walters (2024b): "Race and the mismeasure of school quality," *American Economic Review: Insights*, 6, 20–37.

Angrist, J., P. Hull, and C. Walters (2023): "Chapter 1 - Methods for measuring school effectiveness," in *Handbook of the Economics of Education*, ed. by E. A. Hanushek, S. Machin, and L. Woessmann, Elsevier, vol. 7, 1–60.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2016): "Interpreting tests of school VAM validity," *American Economic Review: Papers & Proceedings*, 106, 388–392.

——— (2017): "Leveraging lotteries for school value-added: testing and estimation," *Quarterly Journal of Economics*, 132, 871–919.

Angrist, J. D., P. A. Pathak., and C. R. Walters (2013): "Explaining charter school effectiveness," *American Economic Journal: Applied Economics*, 5, 1–27.

Arceo-Gomez, E. O. and R. M. Campos-Vasquez (2014): "Race and marriage in the labor market: a discrimination correspondence study in a developing country," *American Economic Review: Papers & Proceedings*, 104, 376–380.

Armstrong, T. B., M. Kolesár, and M. Plagborg-Møller (2022): "Robust empirical Bayes confidence intervals," *Econometrica*, 90, 2567–2602.

Athey, S. and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11, 685–725.

Avivi, H. (2024): "One land, many promises: assessing the consequences of unequal childhood location effects," Working paper.

Bartlett, M. S. (1936): "The square root transformation in analysis of variance," *Supplement to the Journal of the Royal Statistical Society*, 3, 68–78.

Benjamini, Y. and Y. Hochberg (1995): "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, 57, 289–300.

Bergman, P. and M. J. Hill (2018): "The effects of making performance information public: Regression discontinuity evidence from Los Angeles teachers," *Economics of Education Review*, 66, 104–113.

Bertrand, M. and S. Mullainathan (2004): "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, 94, 991–1013.

Beuermann, D. W., C. K. Jackson, L. Navarro-Sola, and F. Pardo (2022): "What is a good school, and can parents tell? Evidence on the multidimensionality of school output," *The Review of Economic Studies*, 90, 65–101.

Bock, M. E. (1975): "Minimax estimators of the mean of a multivariate normal distribution," *The Annals of Statistics*, 3, 209 – 218.

Bonhomme, S. and A. Denis (2024): "Estimating heterogeneous effects: applications to labor economics," Working paper.

Bonhomme, S., K. Holzheu, T. Lamadon, E. Manresa, M. Mogstad, and B. Setzler (2023): "How much should we trust estimates of firm effects and worker sorting?" *Journal of Labor Economics*, 41, 291–322.

Bonhomme, S., T. Lamadon, and E. Manresa (2019): "A distributional framework for matched employer employee data," *Econometrica*, 87, 699–739.

Brown, L. D. (2008): "In-season prediction of batting averages: A field test of empirical Bayes and Bayes methodologies," *The Annals of Applied Statistics*, 2, 113–152.

Card, D. (1999): "The causal effect of education on earnings," Elsevier, vol. 3 of *Handbook of Labor Economics*, 1801–1863.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018): "Firms and labor market inequality: evidence and some theory," *Journal of Labor Economics*, 36, S13–S70.

Card, D., A. R. Cardoso, and P. Kline (2015): "Bargaining, sorting, and the gender wage gap: quantifying the impact of firms on the relative pay of women," *The Quarterly Journal of Economics*, 131, 633–686.

Card, D., J. Heining, and P. Kline (2013): "Workplace heterogeneity and the rise of West German wage inequality," *Quarterly Journal of Economics*, 128, 967–1015.

Chamberlain, G. (1982): "Multivariate regression models for panel data," *Journal of Econometrics*, 18, 5–46.

Chan, D. C., M. Gentzkow, and C. Yu (2022): "Selection with variation in diagnostic skill: evidence from radiologists," *The Quarterly Journal of Economics*, 137, 729–783.

Chandra, A., M. Dalton, and D. O. Staiger (2023): "Are hospital quality indicators causal?" NBER working paper no. 31789.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016): "Health care exceptionalism? Performance and allocation in the US health care sector," *American Economic Review*, 106, 2110–44.

Chen, J. (2023): "Empirical Bayes when estimation precision predicts parameters," ArXiv working paper 2212.14444.

Chetty, R., D. Deming, and J. N. Friedman (2023): "Diversifying society's leaders? The determinants and causal effects of admission to highly selective private colleges," NBER working paper no. 31492.

Chetty, R., J. N. Friedman, N. Hendren, M. R. Jones, and S. R. Porter (2018): "The opportunity atlas: mapping the childhood roots of social mobility," Working paper.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014a): "Measuring the impact of teachers I: evaluating bias in teacher value-added estimates," *American Economic Review*, 104, 2593–2563.

——— (2014b): "Measuring the impact of teachers II: teacher value-added and student outcomes in adulthood," *American Economic Review*, 104, 2633–2679.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2017): "Mobility report cards: the role of colleges in intergenerational mobility," The Equality of Opportunity Project, January.

Chetty, R. and N. Hendren (2018): "Impacts of neighborhoods on intergenerational mobility II: county-level estimates," *Quarterly Journal of Economics*, 133, 1163–1228.

Condorcet, M. d. (1785): "Essay on the application of analysis to the probability of majority decisions," Paris: Imprimerie Royale.

Dale, S. B. and A. B. Krueger (2002): "Estimating the payoff to attending a more selective college: an application of selection on observables and unobservables," *Quarterly Journal of Economics*, 117, 1491–1527.

——— (2014): "Estimating the effects of college characteristics over the career using administrative earnings data," *Journal of Human Resources*, 49, 323–358.

Dobbie, W. and R. G. Fryer (2013): "Getting beneath the veil of effective schools: evidence from New York City," *American Economic Journal: Applied Economics*, 5, 28–60.

Efron, B. (2012): *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1, Cambridge University Press.

——— (2016): "Empirical Bayes deconvolution estimates," *Biometrika*, 103, 1–20.

Efron, B. and C. Morris (1973a): "Combining possibly related estimation problems," *Journal of the Royal Statistical Society. Series B (Methodological)*, 35, 379–421.

——— (1973b): "Stein's estimation rule and its competitors – an empirical Bayes approach," *Journal of the American Statistical Association*, 68, 117–130.

——— (1975): "Data analysis using Stein's estimator and its generalizations," *Journal of the American Statistical Association*, 70, 311–319.

Efron, B. and R. Tibshirani (1996): "Using specially designed exponential families for density estimation," *The Annals of Statistics*, 24, 2431 – 2461.

Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001): "Empirical Bayes analysis of a microarray experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Einav, L., A. Finkelstein, and N. Mahoney (2022): "Producing health: measuring value added of nursing homes," NBER working paper no. 30228.

Evdokimov, K. and H. White (2012): "Some extensions of a lemma of kotlarski," *Econometric Theory*, 28, 925–932.

Fenizia, A. (2022): "Managers and productivity in the public sector," *Econometrica*, 90, 1063–1084.

Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging judge fixed effects," *American Economic Review*, 113, 253–77.

Gilraine, M., J. Gu, and R. McMillan (2020): "A new method for estimating teacher value-added," NBER working paper no. 27094.

Goldhaber, D., B. Gross, and D. Player (2011): "Teacher career paths, teacher quality, and persistence in the classroom: are public schools keeping their best?" *Journal of Policy Analysis and Management*, 30, 57–87.

Goncalves, F. and S. Mello (2021): "A few bad apples? Racial bias in policing," *American Economic Review*, 111, 1406–41.

Gruetter, M. and R. Lalive (2009): "The importance of firms in wage determination," *Labour Economics*, 16, 149–160.

Gu, J. and R. Koenker (2016): "On a problem of Robbins," *International Statistical Review / Revue Internationale de Statistique*, 84, 224–244.

——— (2017): "Unobserved heterogeneity in income dynamics: an empirical Bayes perspective," *Journal of Business & Economic Statistics*, 35, 1–16.

——— (2022): "Ranking and selection from pairwise comparisons: empirical Bayes methods for citation analysis," *AEA Papers and Proceedings*, 112, 624–29.

——— (2023a): "GLVmix: NPMLE of Gaussian location-scale mixture model," https://rdrr.io/cran/REBayes/src/R/GLVmix.R.

——— (2023b): "Invidious comparisons: ranking and selection as compound decisions," *Econometrica*, 91, 1–41.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251–1271.

Heckman, J. J. and B. Singer (1984): "A method for minimizing the impact of distributional assumptions in econometric models for duration data," *Econometrica*, 52, 271–320.

Holland, P. W. (1973): "Covariance stabilizing transformations," *The Annals of Statistics*, 1, 84 – 92.

Jackson, C. K., S. C. Porter, J. Q. Easton, A. Blanchard, and S. Kiguel (2020): "School effects on socioemotional development, school-based arrests, and educational attainment," *American Economic Review: Insights*, 2, 491–508.

James, W. and C. Stein (1961): "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379.

Jiang, W. (2020): "On general maximum likelihood empirical Bayes estimation of heteroscedastic IID normal means," *Electronic Journal of Statistics*, 14, 2272 – 2297.

Jiang, W. and C.-H. Zhang (2009): "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, 37, 1647 – 1684.

Keane, M. and T. Neal (2023): "Instrument strength in IV estimation and inference: A guide to theory and practice," *Journal of Econometrics*, 235, 1625–1653.

Kiefer, J. and J. Wolfowitz (1956): "Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters," *The Annals of Mathematical Statistics*, 27, 887–906.

Kline, P., E. K. Rose, and C. R. Walters (2022): "Systemic discrimination among large U.S. employers," *The Quarterly Journal of Economics*, 137, 1963–2036.

Kline, P., R. Saggio, and M. Sølvsten (2020): "Leave-out estimation of variance components," *Econometrica*, 88, 1859–1898.

Kline, P. and C. Walters (2021): "Reasonable doubt: experimental detection of job-level employment discrimination," *Econometrica*, 89, 765–792.

Kline, P. M., E. K. Rose, and C. R. Walters (2024): "A discrimination report card," *American Economic Review*, 114, 2472–2525.

Koenker, R. (2020): "Empirical Bayes confidence intervals: an R vinaigrette," Working paper.

Koenker, R. and J. Gu (2017): "REBayes: an R package for empirical Bayes mixture methods," *Journal of Statistical Software*, 82, 1–26.

——— (2024): "Empirical Bayes for the reluctant frequentist," ArXiv working paper 2404.30422.

Koenker, R. and I. Mizera (2014): "Convex optimization, shape constraints, compound decisions, and empirical Bayes rules," *Journal of the American Statistical Association*, 109, 674–685.

Kolstad, J. T. (2013): "Information and quality when motivation is itrinsic: evidence from surgeon report cards," *American Economic Review*, 103, 2875–2910.

Kotlarski, I. (1967): "On characterizing the gamma and the normal distribution." *Pacific Journal of Mathematics*, 20, 69 – 76.

Krueger, A. B. and L. H. Summers (1988): "Efficiency wages and the inter-industry wage structure," *Econometrica*, 56, 259–293.

Lachowska, M., A. Mas, R. Saggio, and S. A. Woodbury (2023a): "Do firm effects drift? Evidence from Washington administrative data," *Journal of Econometrics*, 233, 375–395.

——— (2023b): "Work hours mismatch," NBER working paper no. 31205.

Lai, T. L. and D. Siegmund (2018): "Herbert Robbins, 1915-2001: biographical memoir," National Academy of Sciences.

Lehmann, E. L. and J. P. Romano (2005): "Generalizations of the familywise error rate," *The Annals of Statistics*, 33, 1138–1154.

Lindley, D. V. (1962): "Discussion of Professor Stein's Paper," *Journal of the Royal Statistical Society: Series B (Methodological)*, 24, 285–287.

MacKinnon, J. G. and H. White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325.

Manski, C. F. (2000): "Identification problems and decisions under ambiguity: empirical analysis of treatment response and normative analysis of treatment choice," *Journal of Econometrics*, 95, 415–442.

Mogstad, M., J. P. Romano, A. M. Shaikh, and D. Wilhelm (2023): "Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries," *The Review of Economic Studies*, 91, 476–518.

Morris, C. N. (1983): "Parametric empirical Bayes inference: theory and applications," *Journal of the American Statistical Association*, 78, 47–55.

Mountjoy, J. and B. Hickman (2021): "The returns to college(s): relative value-added and match effects in higher education," NBER working paper no. 29276.

Mullainathan, S. and J. Spiess (2017): "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, 31, 87–106.

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Econometrica*, 46, 69–85.

Narasimhan, B. and B. Efron (2020): "deconvolveR: A G-modeling program for deconvolution and empirical Bayes estimation," *Journal of Statistical Software*, 94, 1–20.

Nunley, J. M., A. Pugh, N. Romero, and R. A. Seals (2015): "Racial discrimination in the labor market for recent college graduates: evidence from a field experiment," *B.E. Journal of Economic Analysis and Policy*, 15, 1093–1125.

Pope, D. G. (2009): "Reacting to rankings: evidence from "America's Best Hospitals"," *Journal of Health Economics*, 28, 1154–1165.

Raudenbush, S., S. Reardon, and T. Nomi (2012): "Statistical analysis for multisite trials using instrumental variables with random coefficients," *Journal of Research on Educational Effectiveness*, 5, 303–332.

Robbins, H. (1950): "A generalization of the method of maximum likelihood: estimating a mixing distribution," *The Annals of Mathematical Statistics*, 21, 314–315.

——— (1951): "Asymptotically subminimax solutions of compound statistical decision problems," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, I, 131–139.

——— (1956): "An empirical Bayes approach to statistics," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1, 157–163.

——— (1964): "The empirical Bayes approach to statistical decision problems," *The Annals of Mathematical Statistics*, 35, 1–20.

Rose, E. K., J. T. Schellenberg, and Y. Shem-Tov (2022): "The effects of teacher quality on adult criminal justice contact," NEBER working paper no. 30274.

Savage, L. J. (1951): "The theory of statistical decision," *Journal of the American Statistical association*, 46, 55–67.

Song, J., D. J. Price, F. Guvenen, N. Bloom, and T. von Wachter (2018): "Firming up inequality," *The Quarterly Journal of Economics*, 134, 1–50.

Staiger, D. O. and J. E. Rockoff (2010): "Searching for effective teachers with imperfect information," *Journal of Economic Perspectives*, 24, 97–118.

Stein, C. (1956): "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," 197–206.

Sterne, J. A. C. and R. M. Narbord (2004): "Funnel plots in meta-analysis," *The Stata Journal*, 4, 127–141.

Storey, J. D. (2002): "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64, 479–498.

——— (2003): "The positive false discovery rate: a Bayesian interpretation and the q-value," *The Annals of Statistics*, 31, 2013–2035.

——— (2015): "qvalue: q-value estimation for false discovery rate control," https://github.com/StoreyLab/qvalue.

Storey, J. D., J. E. Taylor, and D. Siegmund (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, 187–205.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.

Varian, H. R. (2014): "Big data: new tricks for econometrics," *Journal of Economic Perspectives*, 28, 3–28.

Wald, A. (1945): "Statistical decision functions which minimize the maximum risk," *Annals of Mathematics*, 46, 265–280.

Walters, C. R. (2015): "Inputs in the production of early childhood human capital: evidence from Head Start," *American Economic Journal: Applied Economics*, 7, 76–102.

——— (2018): "The demand for effective charter schools," *Journal of Political Economy*, 126.

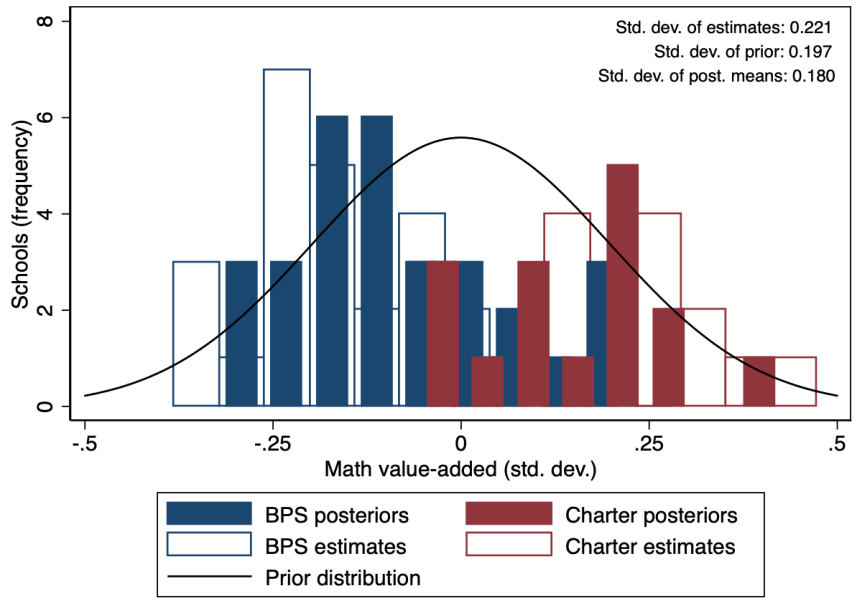
White, H. (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817–838.

Woodcock, S. D. (2008): "Wage differentials in the presence of unobserved worker, firm, and match heterogeneity," *Labour Economics*, 15, 771–793, european Association of Labour Economists 19th annual conference / Firms and Employees.

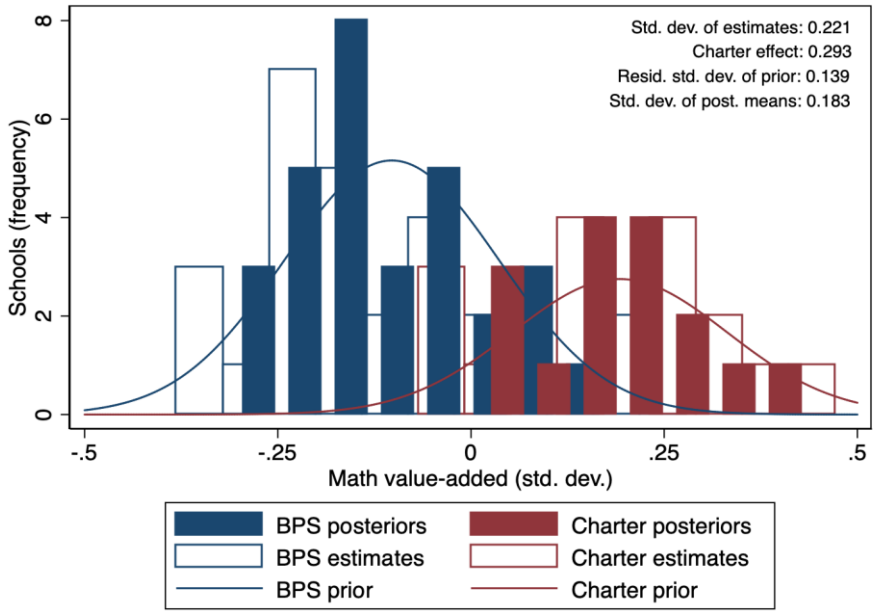
Xie, X., S. C. Kou, and L. D. Brown (2012): "SURE estimates for a heteroscedastic hierarchical model," *Journal of the American Statistical Association*, 107, 1465–1479, pMID: 25301976.

Figure 1: Empirical Bayes estimates of school value-added in Boston

A. Unconditional shrinkage



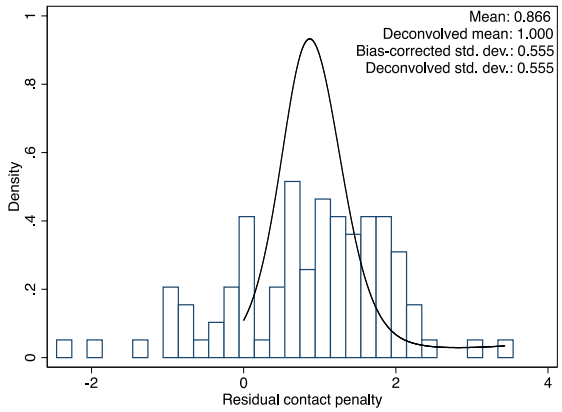
B. Conditional shrinkage



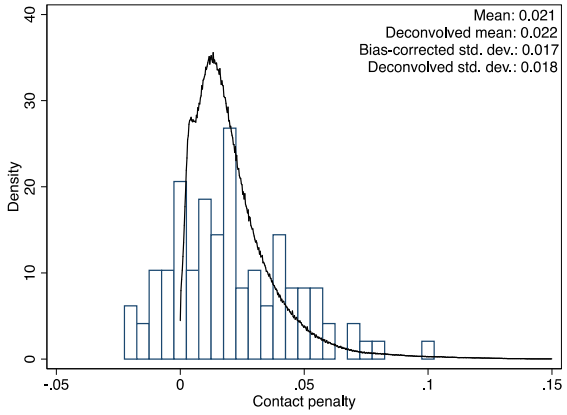
Notes: This figure displays estimates of value-added for 46 Boston middle schools. Outcomes are sixth-grade math scores in 2014 scaled to have mean zero and standard deviation one among all Boston students. Estimates come from regressions of math scores on school indicators with controls for fifth grade math and reading scores, sex, race, subsidized lunch, special education, and English language learner status. School coefficients are centered to have mean zero across all Boston schools. Open histograms plot the distribution of raw value-added estimates with Boston Public Schools (BPS) schools in blue and charter schools in red. Solid curves plot estimated priors based on normal models for the mixing distribution. The prior standard deviation in panel A is calculated by subtracting the average squared standard error from the sample variance of estimates and taking the square root. The charter effect in panel B comes from a regression of value-added estimates on a charter indicator, and the residual standard deviation of the prior is calculated by subtracting the average squared standard error from the raw residual variance then taking the square root. Solid histograms plot linear shrinkage posterior means constructed using the estimated mixing distributions as priors.

Figure 2. Deconvolution estimates of discrimination distributions

A. Distributions of race contact gaps (white - Black)

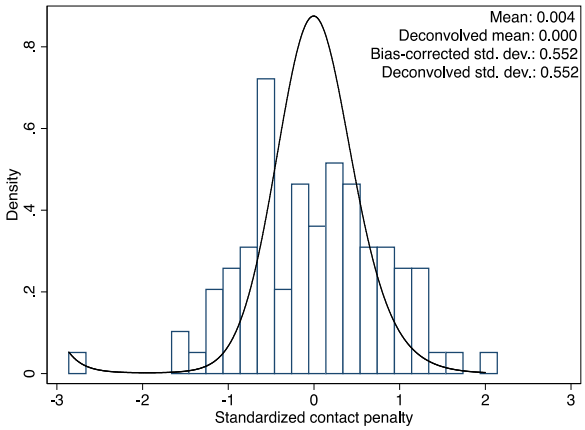


(i). Contact gap residual

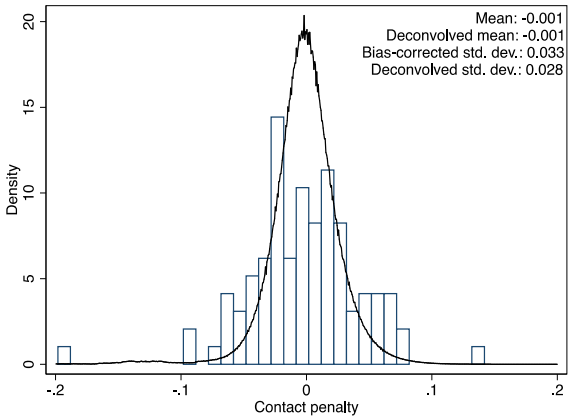


(ii). Contact gap level

B. Distributions of gender contact gaps (male - female)



(i). Contact gap residual

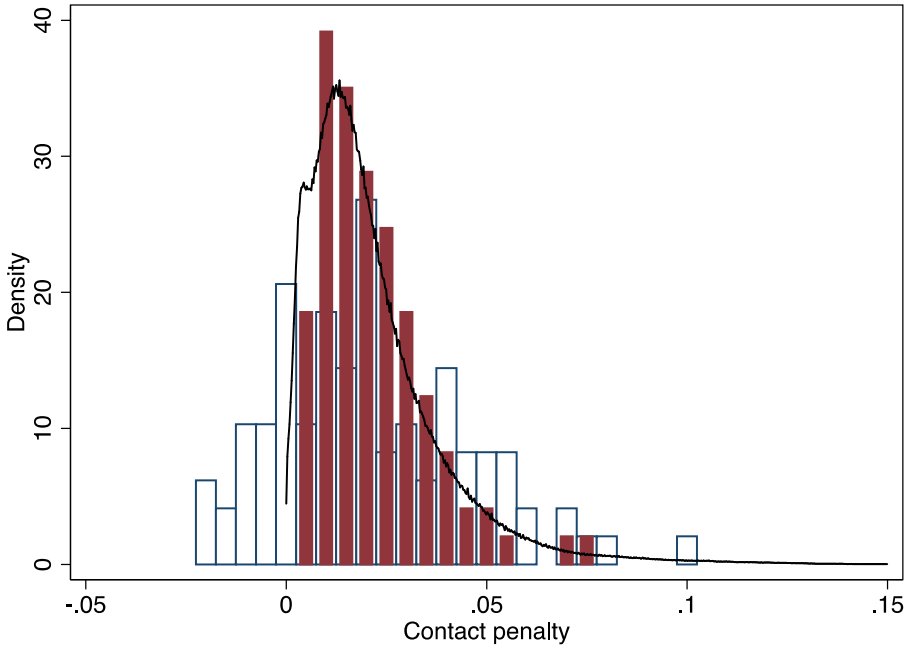


(ii). Contact gap level

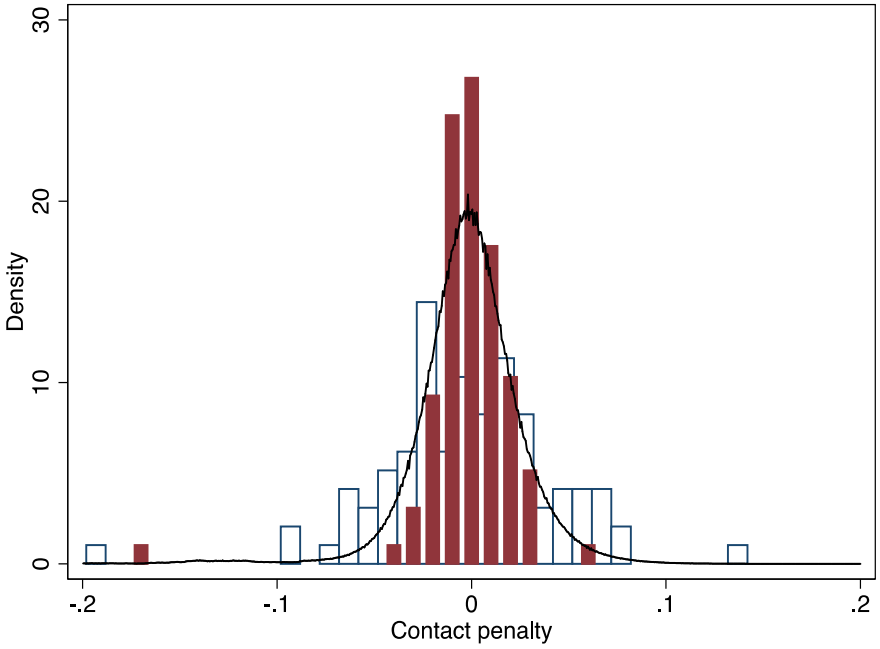
Notes: This figure displays log-spline deconvolution estimates of distributions of differences in contact rates by race and gender across firms. Panel A displays estimated distributions of race gaps (white - Black), and panel B shows estimated distributions of gender gaps (male - female). Blue bars display histograms of observed estimates, and black curves show estimated prior distributions. In each panel, display (i) shows the deconvolved distribution of residuals transformed to eliminate precision-dependence, and display (ii) shows the resulting distribution of contact gap levels, which is constructed by applying a change of variables to the residual distribution and empirical distribution of standard errors. Log-spline estimates come from a penalized maximum likelihood procedure with penalty term calibrated so that bias-corrected and log-spline estimates of residual standard deviations match.

Figure 3. Empirical Bayes estimates of firm-level discrimination

A. Race contact gaps (white - Black)



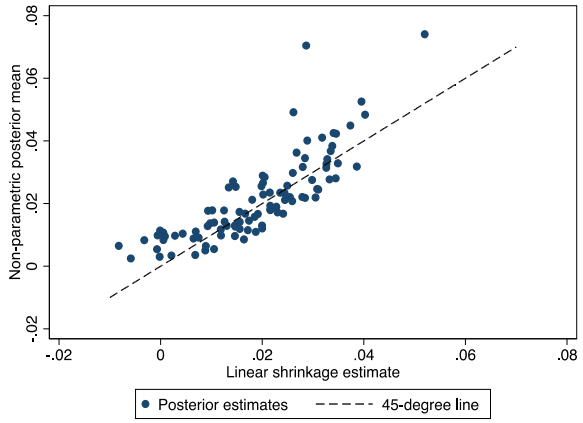
B. Gender contact gaps (male - female)



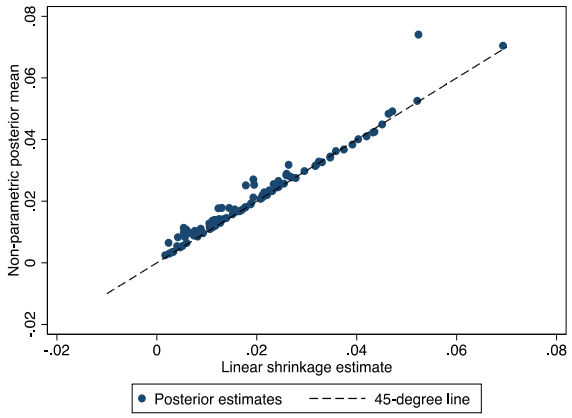
Notes: This figure displays distributions of discrimination estimates for 97 US employers. Blue bars show histograms of unbiased contact gap estimates between applicants with distinctively-white and distinctively-Black names (panel A) or distinctively-male and distinctively-female names (panel B). Black curves show log-spline deconvolution estimates of discrimination distributions. Red bars show non-parametric posterior means that use the log-spline estimates as priors.

Figure 4. Non-parametric posterior mean and linear shrinkage estimates of firm-level discrimination

A. Posterior mean race contact gaps (white - Black)

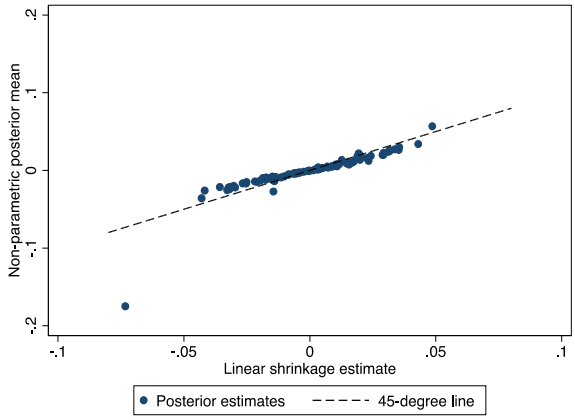


(i). Linear shrinkage ignoring precision-dependence

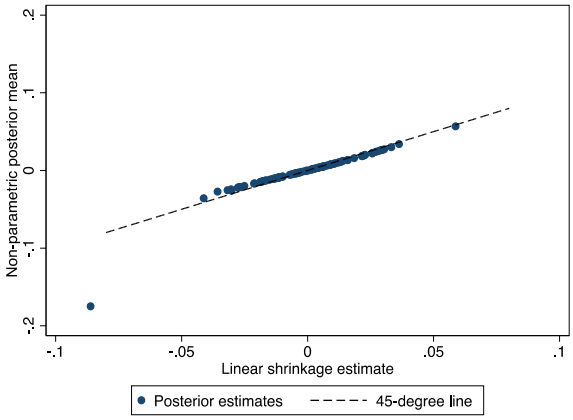


(ii). Linear shrinkage accounting for precision-dependence

B. Posterior mean gender contact gaps (male - female)



(i). Linear shrinkage ignoring precision-dependence

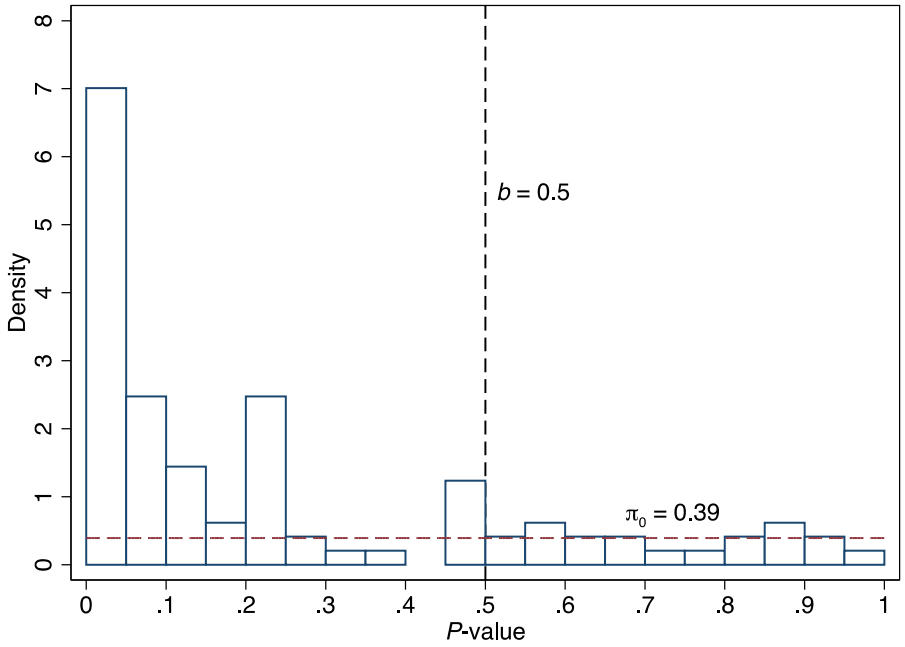


(ii). Linear shrinkage accounting for precision-dependence

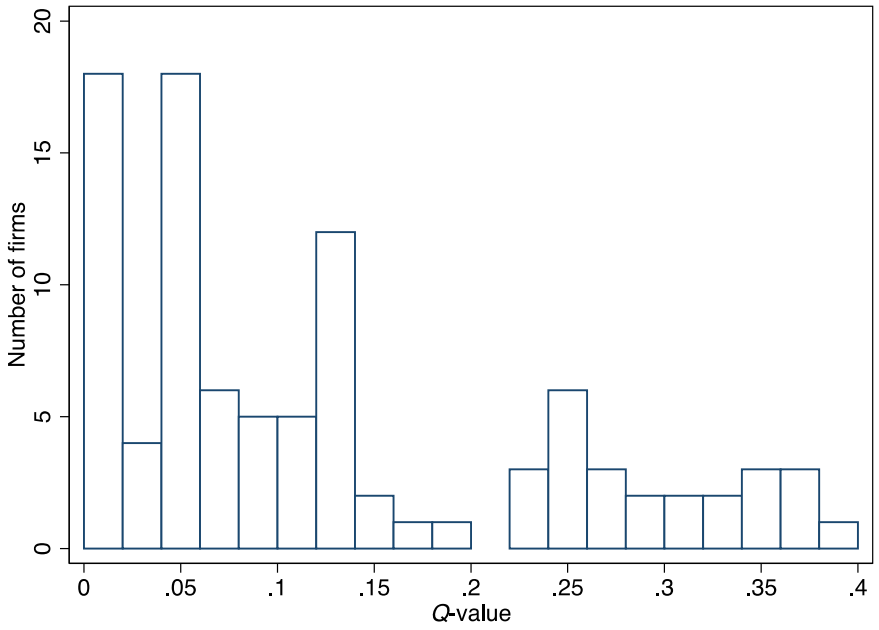
Notes: This figure compares non-parametric and linear shrinkage posterior mean estimates of firm-specific discrimination parameters. Panel A displays posterior mean race gaps (white - Black), and panel B shows posterior mean gender gaps (male - female). Non-parametric posterior means use the deconvolved distributions from Figure 2 as priors. Linear shrinkage posterior means are precision-weighted averages of a firm's unbiased estimate and an estimated prior mean. In each panel, display (i) shows linear shrinkage estimates assuming effect sizes are independent of standard errors, while display (ii) incorporates precision-dependence by applying linear shrinkage to residuals from models relating effect sizes to standard errors, then transforming the resulting posterior residuals to produce posterior contact gaps. Dashed lines are 45-degree lines.

Figure 5. Empirical Bayes multiple-testing analysis of firm-level discrimination

A. P-values for no discrimination against distinctively-Black names

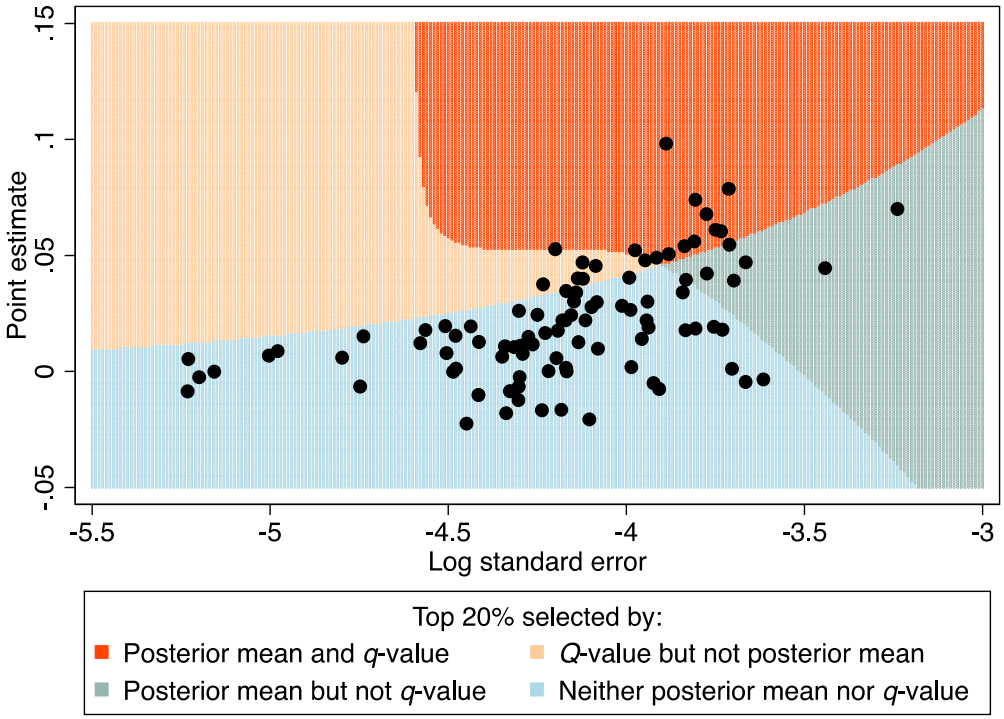


B. Q-values for no discrimination against distinctively-Black names



Notes: This figure displays the results of an empirical Bayes multiple testing analysis of discrimination against distinctively-Black names for 97 firms. Panel A shows p-values from one-tailed z-tests. The black vertical line indicates a threshold of $b = 0.5$ used to bound the share of true nulls, and the red horizontal line displays the resulting estimated bound. Panel B shows a histogram of q-values constructed based on the bound from panel A and the empirical distribution function of p-values.

Figure 6. Selection frontiers for empirical Bayes decision rules



Notes: This figure contrasts empirical Bayes decision rules that select firms based on posterior means and q-values for discrimination against distinctively-Black names. Posterior means use log-spline deconvolution estimates as priors, allowing for dependence between effect sizes and precision. Q-values come from a multiple testing analysis that bounds the prior share of firms that do not discriminate. Decision rules select the 20% of firms with most extreme discrimination estimates based on either the posterior mean or the q-value. The red region shows combinations of point estimates and standard errors that are selected by both posterior mean and q-value decision rules. The yellow region shows combinations selected by a q-value decision rule but not a posterior mean decision rule. The green region shows combinations selected by a posterior mean decision rule but not a q-value decision rule. The blue region shows combinations selected by neither decision rule. Black points indicate the point estimates and standard errors for the 97 firms in the experiment.

Table 1. Variation in race and gender contact gaps across firms

	Race gaps (white - Black)		Gender gaps (male - female)	
	Estimate	Std. err.	Estimate	Std. err.
	(1)	(2)	(3)	(4)
Mean	0.021	0.002	-0.001	0.003
Std. devs.: Uncorrected	0.024	0.002	0.042	0.003
Bias-corrected	0.017	0.003	0.031	0.005
Number of firms	97		97	

Notes: This table reports estimated means and standard deviations of race and gender contact gaps across 97 large US employers. Estimates come from OLS regressions of a callback indicator on an intercept and an indicator for a distinctively-white or distinctively-male name, separately for each firm. Columns (1) and (2) show results for race gaps (white names minus Black names), while columns (3) and (4) display results for gender gaps (male names minus female names). Standard errors for firm-specific OLS coefficients are calculated with a job-clustered covariance matrix. Mean estimates are averages of firm-specific gaps. The uncorrected standard deviation is the square root of the sample variance of firm gap estimates. Bias corrected standard deviations subtract the average squared standard error before taking the square root. Standard errors in columns (2) and (4) come from a job-clustered weighted bootstrap procedure with 1,000 iterations. Each bootstrap iteration draws an *iid* exponential weight for each job and reweights the regressions used to produce firm-specific gaps and standard errors.

Table 2: Empirical Bayes estimates of discrimination against distinctively-Black names

Firm name	Estimate (1)	Std. err. (2)	P-value (3)	Q-value (4)	Non-par. post. mean (5)	Linear shrinkage (6)	Bottom 20% q-value (7)	Top 20% post. mean (8)
Genuine Parts (Napa Auto)	0.098	0.020	0.000	0.000	0.074	0.052	Yes	Yes
AutoNation	0.053	0.015	0.000	0.004	0.032	0.039	Yes	Yes
Advance Auto Parts	0.074	0.022	0.000	0.006	0.048	0.040	Yes	Yes
O'Reilly Automotive	0.079	0.024	0.001	0.006	0.053	0.040	Yes	Yes
Rite Aid	0.047	0.016	0.002	0.012	0.028	0.034	Yes	
Ascena (Ann Taylor / Loft)	0.068	0.023	0.002	0.012	0.045	0.037	Yes	Yes
Gap	0.052	0.019	0.003	0.015	0.033	0.035	Yes	Yes
Dick's	0.051	0.021	0.007	0.015	0.034	0.033	Yes	Yes
Murphy USA	0.049	0.020	0.007	0.016	0.033	0.033	Yes	Yes
Dillard's	0.045	0.017	0.003	0.016	0.028	0.033	Yes	
Tractor Supply	0.040	0.016	0.007	0.016	0.025	0.031	Yes	
AutoZone	0.054	0.022	0.006	0.017	0.037	0.033	Yes	Yes
TJX	0.048	0.019	0.007	0.017	0.031	0.033	Yes	
Bed Bath & Beyond	0.040	0.016	0.006	0.018	0.024	0.031	Yes	
Pizza Hut	0.056	0.022	0.006	0.018	0.038	0.034	Yes	Yes
VFC (North Face / Vans)	0.038	0.014	0.005	0.018	0.022	0.031	Yes	
Pilot Flying J	0.060	0.024	0.006	0.020	0.043	0.034	Yes	Yes
Walgreens	0.061	0.024	0.005	0.020	0.042	0.035	Yes	Yes
Universal Health	0.055	0.024	0.013	0.024	0.041	0.032	Yes	Yes
Sherwin-Williams	0.035	0.015	0.012	0.025	0.022	0.028		
Bath & Body Works	0.040	0.018	0.014	0.026	0.028	0.030		
Olive Garden	0.034	0.016	0.016	0.028	0.022	0.028		
J.C. Penney	0.030	0.016	0.028	0.045	0.021	0.026		
Nationwide	0.026	0.014	0.027	0.045	0.017	0.024		
Dollar General	0.039	0.022	0.034	0.048	0.032	0.028		Yes
Publix	0.047	0.026	0.033	0.049	0.040	0.029		Yes
Marriott	0.030	0.017	0.038	0.049	0.022	0.025		
Republic Services	0.020	0.011	0.038	0.050	0.012	0.020		
Walmart	0.070	0.039	0.037	0.050	0.070	0.029		Yes
US Foods	0.024	0.014	0.044	0.050	0.017	0.023		
Victoria's Secret	0.042	0.023	0.033	0.050	0.034	0.028		Yes
Goodyear	0.015	0.009	0.042	0.052	0.009	0.016		
Hertz	0.018	0.010	0.044	0.052	0.011	0.019		
Builders FirstSource	0.028	0.017	0.047	0.053	0.021	0.024		
Aramark	0.019	0.012	0.051	0.055	0.013	0.020		
Tyson Foods	0.024	0.016	0.061	0.058	0.019	0.023		
Comcast	0.039	0.025	0.057	0.059	0.036	0.027		Yes
Best Buy	0.028	0.018	0.059	0.059	0.023	0.024		
Cardinal Health	0.034	0.021	0.056	0.059	0.030	0.026		
Dollar Tree	0.030	0.019	0.061	0.059	0.026	0.025		
UGI	0.022	0.015	0.077	0.068	0.018	0.022		
LKQ Auto	0.027	0.019	0.076	0.069	0.023	0.024		
KFC	0.022	0.015	0.075	0.070	0.018	0.022		
Dean Foods	0.044	0.032	0.082	0.071	0.049	0.026		Yes
Costco	0.022	0.016	0.089	0.073	0.019	0.022		
CVS Health	0.015	0.011	0.088	0.074	0.011	0.017		
PepsiCo	0.009	0.007	0.102	0.083	0.005	0.011		
Macy's	0.012	0.010	0.118	0.093	0.010	0.015		
Jones Lang LaSalle	0.018	0.015	0.123	0.095	0.017	0.019		
Sears (incl. repair / auto)	0.017	0.015	0.129	0.096	0.016	0.019		
Cintas	0.022	0.019	0.128	0.097	0.024	0.021		
Kohl's	0.015	0.014	0.143	0.105	0.015	0.017		
Edward Jones	0.013	0.012	0.147	0.106	0.012	0.016		
CBRE	0.007	0.007	0.155	0.109	0.005	0.009		
AECOM	0.005	0.005	0.159	0.110	0.004	0.007		
DISH	0.019	0.019	0.166	0.113	0.023	0.020		
Nordstrom	0.011	0.014	0.208	0.127	0.013	0.015		
Estee Lauder	0.018	0.022	0.206	0.128	0.026	0.020		
XPO Logistics	0.010	0.013	0.217	0.129	0.013	0.015		
Performance Food Group	0.019	0.023	0.206	0.130	0.028	0.020		
International Paper	0.013	0.016	0.216	0.130	0.017	0.017		
Honeywell	0.011	0.013	0.204	0.131	0.013	0.015		
Geico	0.018	0.024	0.227	0.133	0.029	0.020		
GameStop	0.008	0.011	0.238	0.133	0.010	0.012		
US Bank	0.012	0.014	0.204	0.133	0.014	0.016		
Starbucks	0.014	0.019	0.232	0.134	0.021	0.018		
UnitedHealth	0.006	0.008	0.236	0.134	0.006	0.009		
Ross Stores	0.018	0.022	0.203	0.136	0.027	0.020		
AT&T	0.010	0.017	0.281	0.155	0.017	0.016		
Stanley Black & Decker	0.008	0.014	0.290	0.157	0.013	0.013		
Foot Locker	0.006	0.013	0.312	0.167	0.012	0.012		
United Rentals	0.006	0.015	0.352	0.186	0.014	0.013		
Ulta Beauty	0.002	0.015	0.461	0.233	0.014	0.011		
WestRock	0.002	0.019	0.460	0.236	0.018	0.012		
Target	0.001	0.011	0.457	0.238	0.009	0.007		
CarMax	0.001	0.025	0.482	0.241	0.025	0.015		
Home Depot	0.000	0.006	0.508	0.241	0.003	0.002		
Kroger	0.000	0.011	0.506	0.243	0.009	0.006		
Disney (incl. stores)	0.000	0.015	0.500	0.244	0.014	0.010		
Waste Management	0.000	0.015	0.497	0.246	0.013	0.009		
Safeway	-0.003	0.027	0.552	0.259	0.027	0.014		
McLane Company	-0.005	0.026	0.571	0.261	0.025	0.013		
Lab Corp	-0.002	0.014	0.571	0.265	0.011	0.007		
Mondelez	-0.005	0.020	0.601	0.272	0.018	0.010		
Lowe's	-0.008	0.020	0.648	0.290	0.018	0.009		
Sysco	-0.003	0.006	0.675	0.298	0.003	0.000		
FedEx	-0.007	0.014	0.687	0.300	0.010	0.004		
State Farm	-0.009	0.013	0.741	0.320	0.010	0.003		
Ryder System	-0.007	0.009	0.775	0.331	0.005	-0.001		
Charter / Spectrum	-0.010	0.012	0.800	0.338	0.008	0.001		
JPMorgan Chase	-0.012	0.014	0.820	0.342	0.010	0.001		
J.B. Hunt	-0.017	0.015	0.861	0.356	0.011	0.001		
Avis-Budget	-0.017	0.014	0.877	0.358	0.010	-0.001		
Kindred Healthcare	-0.021	0.017	0.895	0.362	0.011	0.000		
Hilton	-0.018	0.013	0.916	0.367	0.008	-0.003		
Quest Diagnostics	-0.009	0.005	0.947	0.375	0.002	-0.006		
Dr Pepper	-0.023	0.012	0.973	0.381	0.007	-0.008		

Notes: This table reports the results of an empirical Bayes analysis of discrimination against distinctively-Black names among 97 employers. Firms are ordered by q-values from a multiple testing procedure. Column (1) shows each firm's estimated difference in contact rates between distinctively-white and distinctively-Black names, and column (2) displays corresponding standard errors. Column (3) shows p-values from one-tailed tests of the null hypothesis of no discrimination against distinctively-Black names, and column (4) shows q-values. Column (5) displays non-parametric posterior means using log-spline deconvolution estimates as priors, incorporating dependence between effect sizes and standard errors. Column (6) shows linear shrinkage posterior means ignoring precision-dependence. Column (7) labels firms in the bottom 20% of the q-values from column (4). Column (8) labels firms in the top 20% of the non-parametric posterior means from column (5).