

REASONABLE UNCERTAINTY: CONFIDENCE INTERVALS IN EMPIRICAL BAYES DISCRIMINATION DETECTION

JIAYING GU, NIKOLAOS IGNATIADIS, AND AZEEM M. SHAIKH

ABSTRACT. We revisit empirical Bayes discrimination detection, focusing on uncertainty arising from both partial identification and sampling variability. While prior work has mostly focused on partial identification, we find that some empirical findings are not robust to sampling uncertainty. To better connect statistical evidence to the magnitude of real-world discriminatory behavior, we propose a counterfactual odds-ratio estimand with attractive properties and interpretation. Our analysis reveals the importance of careful attention to uncertainty quantification and downstream goals in empirical Bayes analyses.

1. INTRODUCTION

Empirical Bayes (Robbins, 1956; Efron, 2019) methods are increasingly popular in applied research. Prominent examples covering a diverse array of applications include studies by Rozema and Schanzenbach (2019) on police behavior, Wernerfelt, Tuchman, Shapiro, and Moakler (2022) on advertising treatment effects, Gu and Koenker (2022) on journal ratings, Metcalfe, Sollaci, and Syverson (2023) on managerial productivity effects, and Coey and Hung (2022) on online controlled experiments. The growing importance of empirical Bayes methods is further highlighted by Walters (2024), who surveys applications in labor economics. A distinguishing feature of these and other empirical Bayes analyses is that they rarely include uncertainty quantification for posterior estimands.

In this paper, we revisit this methodological shortcoming in the context of a compelling application of empirical Bayes described in Kline and Walters (2021) to discrimination detection based on correspondence experiments. Their analysis emphasizes the role of uncertainty stemming from partial identification, but, with a few notable exceptions that we describe further below in Section 3, Kline and Walters primarily report point estimates for their posterior estimands without further incorporating sampling uncertainty. Our discussion, by contrast, highlights the way in which partial identification and sampling uncertainty for posterior estimands are intertwined and naturally addressed in concert. In the course of doing so, we draw attention to some new results on confidence intervals for empirical Bayes analyses and introduce some novel methods exploiting inference methods recently developed for other problems.

We apply these methods to data from one of the three correspondence experiments analyzed by Kline and Walters, specifically the study by Arceo-Gomez and Campos-Vazquez (2014) of race and gender discrimination in Mexico City. After doing so, we find that some of the empirical conclusions in Kline and Walters (2021) concerning this data prove substantially more robust than others. In this way, our analysis demonstrates the importance of accounting for sampling uncertainty in empirical Bayes discrimination analysis. We further show that this remains true for an alternative estimand based on an odds ratio that we argue may be preferred relative to the one considered in Kline and Walters (2021) for several different reasons. Through these contributions, we hope to make it routine to report confidence intervals alongside point estimates for

empirical Bayes estimands and, as called for by Imbens (2022), to encourage further research on statistical inference accompanying empirical Bayes analyses.

The remainder of our paper is organized as follows. In Section 2, we first review the key ingredients of an empirical Bayes analysis and then specialize our discussion to the setting in Kline and Walters (2021). In Section 3, we discuss partial identification and sampling uncertainty through the lens of four different optimization problems. This discussion motivates a particular approach to account for uncertainty described in Section 4 based on Ignatiadis and Wager (2022). Other methods to account for uncertainty are described in Section 5, including a novel application of Fang, Santos, Shaikh, and Torgovitsky (2023). Along the way, we apply each of these methods to re-assess some empirical conclusions in Kline and Walters (2021). Finally, in Section 6, we introduce and discuss our alternative estimand, and apply these methods to it as well.

2. SETUP AND NOTATION

A canonical empirical Bayes analysis (Robbins, 1956; Efron, 2019; Ignatiadis and Wager, 2022) consists of three ingredients:

- Data from multiple related units $Z_1, \dots, Z_n \in \mathcal{Z}$ drawn independently based on a known likelihood, $Z_i \sim p(\cdot \mid \theta_i)$, where θ_i is the parameter of interest for the i -th unit.
- A structural distribution G describing the ensemble of parameters θ_i via $\theta_i \sim G$ and $G \in \mathcal{G}$, where \mathcal{G} is a class of distributions.
- An estimand $\theta(G; z)$ that permits an oracle decision maker with knowledge of the ensemble G to make an optimal decision regarding a unit with observed data z .

The empirical Bayesian has no knowledge of the ensemble G , yet seeks to mimic the oracle decision maker by learning from indirect evidence, i.e., by using the observed Z_1, \dots, Z_n to learn about G . The connection between the observed data and the unknown ensemble is established through the marginal density of the Z_i ,

$$(2.1) \quad f_G(z) = \int p(z \mid \theta) dG(\theta).$$

In the setting considered by Kline and Walters (2021),

- Each unit $i = 1, \dots, n$ is a job. The experimenter sends out L fictitious job applications from each of two groups, labeled a and b , and records the number of callbacks for each group, $Z_i = (C_{ai}, C_{bi}) \in \mathcal{Z} = \{0, \dots, L\}^2$. The likelihood is modeled as a bivariate binomial, i.e., for $z = (c_a, c_b)$,

$$(2.2) \quad p(z \mid \theta) = \binom{L}{c_a} \binom{L}{c_b} p_a^{c_a} (1 - p_a)^{L - c_a} p_b^{c_b} (1 - p_b)^{L - c_b},$$

where $\theta = (p_a, p_b)$ are the callback probabilities for groups a and b , respectively. For such discrete data, the marginal density f_G in (2.1) is a probability mass function (i.e., a density with respect to the counting measure).

- The structural distribution G is a distribution over the unit cube $[0, 1]^2$ and \mathcal{G} is the class of all distributions over $[0, 1]^2$, i.e., no further restrictions are imposed on G .
- The estimand of interest is the posterior probability that a job with callback pattern z discriminates against group b (i.e., favors group a over b),

$$(2.3) \quad \theta^{\text{discr}}(G; z) := \mathbb{P}_G[p_a > p_b \mid Z = z].$$

Kline and Walters demonstrate using data from three different correspondence experiments how empirical Bayes provides a principled way to detect discrimination patterns across jobs in this type of setting. There are, however, two important sources of uncertainty in such an analysis. First, in some empirical Bayes problems, such as the bivariate binomial model in (2.2) described above, even if we had precise knowledge of f_G , we could not recover G uniquely. In other words, G is only partially identified. Second, in practice, we do not know f_G either and must estimate it from data.

Before proceeding, we note that we confine our analysis below to data from one of the three correspondence experiments analyzed by Kline and Walters, specifically the study by Arceo-Gomez and Campos-Vazquez (2014) of gender discrimination, but the same considerations apply equally well to the other two correspondence experiments they analyze, namely Bertrand and Mullainathan (2004) and Nunley, Pugh, Romero, and Seals (2015).

3. ON PARTIAL IDENTIFICATION AND SHAPE-CONSTRAINED GMM

In this section, we discuss four different optimization problems that will help us not only explain the way in which Kline and Walters address the partial identification issue, but also how sampling variability and partial identification are intertwined. The optimization problems each seek to minimize the estimand $\theta(\tilde{G}; z)$ over all distributions \tilde{G} , but are distinguished by different additional constraints.¹ To solve each optimization problem, we discretize \tilde{G} on a two-dimensional grid with K^2 points; we defer a more detailed discussion of computational issues to Section 4.1.

$$(3.1) \quad \begin{array}{ll} \text{minimize} & \theta(\tilde{G}; z) \quad \text{subject to one of:} \\ \tilde{G} \in \mathcal{G} & \end{array} \quad \begin{array}{llll} (i) & f_{\tilde{G}} = f_G, & (ii) & f_{\tilde{G}} = \bar{f}, & (iii) & f_{\tilde{G}} = \bar{f}^{\text{proj}}, & (iv) & J_n(f_{\tilde{G}}, \bar{f}) \leq \kappa. \end{array}$$

Below, we will explain each of these optimization problems in turn and their constraints, defining all required quantities along the way.

Optimization problem (i) represents an idealized benchmark: if we knew the true marginal density f_G , what would be the smallest possible value of $\theta(\tilde{G}; z)$ among all \tilde{G} that are consistent with this density? Though one could also maximize, we focus on the minimum as, for the discrimination probability $\theta^{\text{discr}}(G; z)$, it represents the most conservative value that is compatible with the true marginal density. Optimization problem (i) captures the fundamental partial identification challenge. We next turn to problems (ii)–(iv) that also capture issues that stem from the fact that f_G is not known and must be estimated.

In optimization problems (ii) and (iii) the true density f_G is replaced by an estimate. Problem (ii) uses the empirical frequencies $\bar{f}(z) = \sum_{i=1}^n \mathbf{1}(Z_i = z)/n$, which provide a natural estimate for discrete data.² Kline and Walters pursue precisely this approach to compute estimates of lower bounds in their application to the Bertrand and Mullainathan (2004) experiment. Problem (ii) may, however, be infeasible: there may be no distribution \tilde{G} whose implied marginal density exactly matches \bar{f} . This situation can occur due to sampling variability in \bar{f} , or due to misspecification of the bivariate binomial model. Indeed, infeasibility occurs in two of the three empirical examples in Kline and Walters (2021) and is a well understood phenomenon in the related univariate binomial problem (Wood, 1999).

¹The reader should keep the estimand (2.3) in mind, but the discussion applies to any estimand $\theta(G; z)$.

²For continuous data, estimation of f_G would require more sophisticated density estimation techniques.

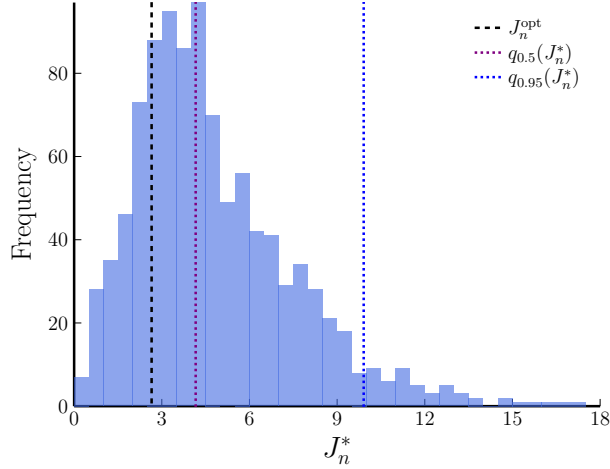


FIGURE 1. CNS bootstrap distribution of J_n^{opt} for the AGCV dataset. Based on 1,000 replicates.

To address infeasibility, Kline and Walters introduce the following generalized method of moments (GMM) problem:

$$(3.2) \quad \underset{\bar{G} \in \mathcal{G}}{\text{minimize}} \quad J_n(f_{\bar{G}}, \bar{f}), \quad J_n(f, \bar{f}) := n(f - \bar{f})^\top \widehat{W}(f - \bar{f}),$$

where \widehat{W} is a weighting matrix computed in a first-stage GMM step. Let G^{proj} be the solution to (3.2), $\bar{f}^{\text{proj}} = f_{G^{\text{proj}}}$ the implied marginal density, and $J_n^{\text{opt}} = J_n(\bar{f}^{\text{proj}}, \bar{f})$ the optimal value of (3.2). Optimization problem (iii) then replaces \bar{f} in the constraint for problem (ii) with \bar{f}^{proj} . Unlike problem (ii), problem (iii) is always feasible by construction, but still ignores sampling variability.

Following common practice for empirical Bayes analyses (as discussed in the introduction), Kline and Walters only report point estimates of their lower bounds for the posterior discrimination estimand in (2.3) without further incorporating sampling uncertainty, that is, they report the objective value of optimization problem 3.1(iii) (or (ii) when feasible). We emphasize, however, that Kline and Walters are aware of sampling uncertainty more generally and address it in some other parts of their analysis. They propose, for example, a shape-constrained bootstrap scheme, following Chernozhukov, Newey, and Santos (2023) (CNS) for goodness-of-fit testing based on the distribution of J_n^{opt} (the minimum value of the GMM statistic). Beyond goodness-of-fit testing, Kline and Walters also adapt the bootstrap scheme of CNS to test null hypotheses such as $\mathbb{P}_G[p_a \neq p_b] = 0$ or $\mathbb{P}_G[p_a > p_b] = 0$.

Figure 1 shows the bootstrap distribution³ of J_n^{opt} for the study of gender discrimination by Arceo-Gomez and Campos-Vazquez (2014) (AGCV), where $J_n^{\text{opt}} \approx 2.65$. Under the null hypothesis that the true probabilities are consistent with the bivariate binomial mixture model, the bootstrap distribution suggests that much larger values than J_n^{opt} can be realized under the null—its 95% quantile is 9.9.

Building on their bootstrap analysis, we propose optimization problem (iv) to illustrate how uncertainty affects their bounds by allowing all distributions with $J_n(f_{\bar{G}}, \bar{f}) \leq \kappa$ for some choice of $\kappa > 0$. For $\kappa < J_n^{\text{opt}}$, the problem is infeasible. When $\kappa = J_n^{\text{opt}}$ and assuming uniqueness of the minimizer, problems (iii) and (iv) yield identical optimal values. For $\kappa > J_n^{\text{opt}}$, problem (iv)’s optimal value can be strictly smaller, reflecting the incorporated additional uncertainty.

³We generate the bootstrap samples by directly rerunning the reproduction code of Kline and Walters (2021).

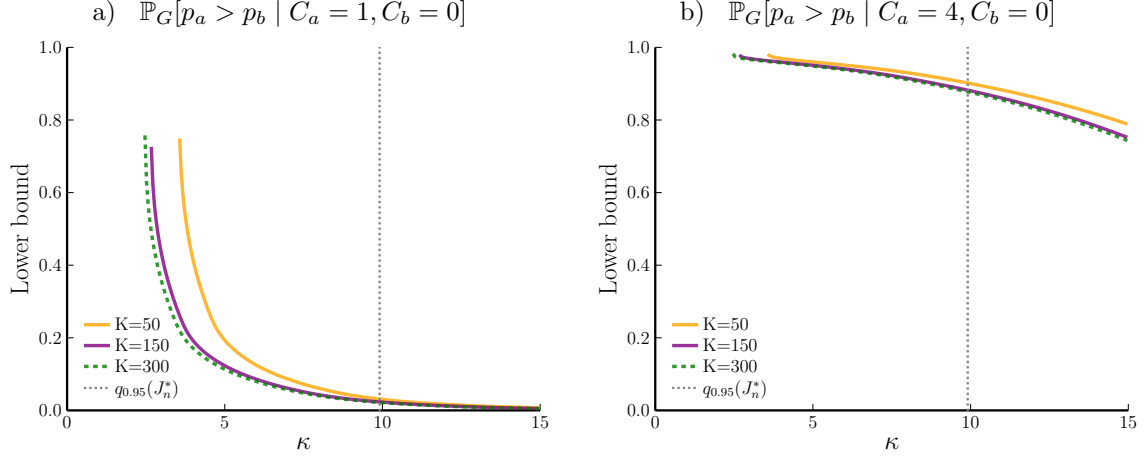
FIGURE 2. Lower bounds as a function of the slack κ .

Figure 2 shows the results of solving optimization problem (iv) for two callback patterns in the AGCV dataset and for different values of κ . In panel a) for $\theta^{\text{discr}}(G, (1, 0))$ and in panel b) for $\theta^{\text{discr}}(G, (4, 0))$. For each pattern, we present results using three different discretization strategies corresponding to $K = 50, 150$, or 300 . All lower bound curves begin at $\kappa = J_n^{\text{opt}}$, with finer discretizations yielding slightly smaller values of J_n^{opt} . While discretization choices substantially impact the lower bounds when κ is close to J_n^{opt} , these effects become less pronounced as κ increases.

The analysis reveals stark differences in robustness across discrimination estimands for different callback patterns. Kline and Walters’ estimate that “an employer that calls back a single woman and no men has at least a 74% chance of discriminating against men” proves highly sensitive to uncertainty—the lower bound rapidly drops as soon as we relax κ beyond J_n^{opt} , falling to just 2% at $\kappa = 9.9$ (the 95% quantile of the bootstrap distribution of J_n^{opt}). By contrast, their estimate that “at least 97% of the jobs that call back four women and no men are estimated to discriminate against men” demonstrates greater robustness, maintaining a lower bound of 88% even at $\kappa = 9.9$. As explained in Section 4 below, this particular choice of κ is a special case of the F -localization approach of Ignatiadis and Wager (2022) and ensures that the lower bound is a valid 95% lower confidence bound on $\theta(G; z)$. Table 1 records this lower confidence bound as well as lower confidence bounds from two other methods that are described in detail in Section 5. Collectively, these results demonstrate the importance of accounting for sampling uncertainty in empirical Bayes discrimination analysis, as some findings prove substantially more robust than others.

	$(C_a, C_b) = (1, 0) \quad (C_a, C_b) = (4, 0)$	
F -Localization ($\kappa = 9.9$)	0.02	0.88
AMARI	0.01	0.92
FSST	0.01	0.89

TABLE 1. Lower bounds of 95% confidence intervals for $\theta^{\text{discr}}(G; z)$ for $z = (1, 0)$ and $z = (4, 0)$.

4. ON THE PRINCIPLE OF F -LOCALIZATION

As mentioned previously, setting κ equal to the 95% quantile of the bootstrap distribution of J_n^{opt} is a special case of the F -localization approach of Ignatiadis and Wager (2022) and ensures that the lower bound

obtained in this way is a valid lower confidence bound on $\theta(G; z)$. To see why, suppose that we choose a potentially data-driven $\hat{\kappa}$ such that $\mathbb{P}_G[J_n(f_G, \bar{f}) \leq \hat{\kappa}] \geq 1 - \alpha$. Fix an estimand of interest $\theta(G; z)$ and solve optimization problem (3.1)(iv) at $\kappa = \hat{\kappa}$ calling the optimal value $\underline{\theta}(\hat{\kappa})$. It follows that

$$\mathbb{P}_G[\theta(G; z) \geq \underline{\theta}(\hat{\kappa})] \geq \mathbb{P}_G[J_n(f_G, \bar{f}) \leq \hat{\kappa}] \geq 1 - \alpha,$$

where the first inequality follows since on the event $\{J_n(f_G, \bar{f}) \leq \hat{\kappa}\}$, G is a feasible solution in optimization problem (3.1)(iv).

The above construction is a special case of the F -localization principle (Ignatiadis and Wager, 2022). The key idea is to construct a $(1 - \alpha)$ -confidence set of marginal distributions $\mathcal{F}(\alpha)$ such that $\mathbb{P}_G[F_G \in \mathcal{F}(\alpha)] \geq 1 - \alpha$, where F_G denotes the marginal distribution of Z under G , i.e., the distribution with density f_G defined in (2.1). From this confidence set, one can construct confidence intervals for any functional of interest $\theta(G; z)$ by optimizing over all distributions G whose marginals lie in $\mathcal{F}(\alpha)$, similar to the optimization problem in (3.1)(iv).⁴

In this way, the F -localization approach translates statements concerning the uncertainty about the distribution of observables F_G into statements concerning the uncertainty about the latent distribution G and functionals thereof. The projection idea underlying the F -Localization approach traces back to the fundamental ideas of Scheffé (1953) and Anderson (1969). There are several ways of constructing F -localizations. For instance, one generic approach that works for any univariate Z_i is to use the Dvoretzky-Kiefer-Wolfowitz inequality with Massart's (1990) tight constant. In other cases, more specialized and refined constructions can work instead; for instance, Ignatiadis and Wager (2022) construct a F -localization in the Gaussian empirical Bayes problem by considering an L_∞ neighborhood of the marginal density f_G . For discrete problems, such as the one considered by Kline and Walters, methods using a χ^2 -based F -localization were already developed by Lord and Cressie (1975); Lord and Stocking (1976).

An important feature of the F -localization principle is that it can be used to construct confidence intervals for any functional of the distribution G , provided that the resulting optimization problem is tractable. Moreover, all these confidence intervals have simultaneous $1 - \alpha$ coverage: if $F_G \in \mathcal{F}(\alpha)$, which has probability at least $1 - \alpha$ if $\mathcal{F}(\alpha)$ is a valid F -localization, then all the resulting confidence intervals will cover the true value of the functional with at least the desired probability. Simultaneity is a desirable property when the empirical Bayes analysis is highly exploratory as in Kline and Walters: therein the authors consider all kinds of estimands: $\theta^{\text{discr}}(G; z)$ for different callback patterns z ; alternative definitions of discrimination, e.g., $\mathbb{P}_G[p_a \neq p_b \mid Z = z]$ (again, for different z); unconditional discrimination probabilities such as $\mathbb{P}_G[p_a > p_b]$ and so forth. The F -localization principle allows one to construct confidence intervals for all of these estimands with simultaneous coverage.

4.1. Computational issues for F -localization. It is common in empirical Bayes problems to discretize the space of distributions \mathcal{G} (Koenker and Mizera, 2014). For the discrimination detection problem, where \mathcal{G} consists of distributions over $[0, 1]^2$, we introduce a grid $\mathcal{D}_K = \{\theta_\ell : \ell = 1, \dots, K^2\} \subset [0, 1]^2$ (following Kline and Walters, 2021, Appendix C) and represent distributions as $\tilde{G} = \sum_{\ell=1}^{K^2} \pi_\ell \delta_{\theta_\ell}$ where δ_{θ_ℓ} denotes the Dirac measure at $\theta_\ell \in [0, 1]^2$, and the weights π_ℓ satisfy $\pi_\ell \geq 0$ and $\sum_\ell \pi_\ell = 1$. Ideally, one should use as fine a

⁴It is possible that there is no $\tilde{G} \in \mathcal{G}$ such that $F_{\tilde{G}} \in \mathcal{F}(\alpha)$. If $\mathcal{F}(\alpha)$ is a valid F -localization, then this can happen for two reasons: we are on the low probability event that $F_G \notin \mathcal{F}(\alpha)$ or the model is misspecified. Thus the F -Localization approach includes an embedded specification test, similar to e.g., Romano and Shaikh (2008) and Stoye (2009).

grid as computationally feasible. In Figure 2, we use, like Kline and Walters, the above discretization scheme with varying grid sizes corresponding to $K = 50, 150, 300$ to assess sensitivity to discretization.

Following Kline and Walters, let us explain how this discretization turns optimization problems (i)–(iii) into linear programs. Common empirical Bayes estimands, including the discrimination estimand $\theta^{\text{discr}}(G; z)$, take the form,

$$(4.1) \quad \theta^{\text{post}}(G; z) = \mathbb{E}[h(\theta) \mid Z = z] = \frac{\int h(\theta)p(z \mid \theta)dG(\theta)}{f_G(z)},$$

for a function $h(\cdot)$, e.g., $h(\theta) = \mathbb{1}(p_a > p_b)$. For such estimands, after discretization, one can solve optimization problem (3.1)(iii) (and analogously, (i) and (ii)) by linear programming:

$$\underset{\pi \in [0,1]^{K^2}}{\text{minimize}} \quad \sum_{\ell=1}^{K^2} h(\theta_\ell) \frac{p(z \mid \theta_\ell)}{\bar{f}^{\text{proj}}(z)} \pi_\ell \quad \text{s.t.} \quad \sum_{\ell=1}^{K^2} p(z' \mid \theta_\ell) \pi_\ell = \bar{f}^{\text{proj}}(z') \quad \text{for all } z' \in \mathcal{Z}, \quad \sum_{\ell=1}^{K^2} \pi_\ell = 1.$$

Observe that the objective and the constraints are linear in the π_ℓ .⁵ Analogously, after discretization, optimization problem (3.2) is also a convex program that can be solved by second order conic programming (SOCP).⁶

It turns out that we can substantially extend both the class of estimands and the constraints (beyond linear) and still maintain computational tractability that facilitates the construction of F -localization based confidence intervals. Concretely, consider any estimand that may be written as a ratio of linear functionals of G ,

$$(4.2) \quad \theta^{\text{ratio}}(G; z) = \frac{N(G; z)}{D(G; z)},$$

with $N(G; z)$ and $D(G; z)$ linear functionals of G .⁷ Then, the optimization problem (3.1)(iv) can also be solved as a SOCP using techniques from fractional programming (Charnes and Cooper, 1962); see Ignatiadis and Wager (2022) for details. Hence, e.g., computing the lower bounds in Figure 2 for the discrimination estimand $\theta^{\text{discr}}(G; z)$ is computationally fast even for the grid with 300^2 points.

5. INFERENCE METHODS BEYOND F -LOCALIZATION

In some situations, because it achieves simultaneous coverage over all possible empirical Bayes estimands, F -localization may be overly conservative. Some recent innovations permit construction of confidence intervals that have nominal coverage for a specific estimand of interest, and so, in some cases, can be substantially shorter than F -localization intervals. Here, we describe two approaches, both of which account for both sources of uncertainty, partial identification and sampling variability. Discretization considerations for these methods are similar to those for F -localization.

5.1. Affine Minimax Anderson-Rubin Inference (AMARI). This method developed in Ignatiadis and Wager (2022) provides confidence intervals for any ratio estimand of the form in (4.2). The starting point is to test for each c whether $\theta^{\text{ratio}}(G; z) = c$ and to obtain a confidence interval by inversion. By an

⁵Note that \bar{f}^{proj} is computed in a first step in a separate optimization problem and is treated as fixed in the linear program; by doing so, the ratio objective becomes linear in the optimization variables. The fractional programming techniques we describe below allow directly solving optimization problems with a ratio objective.

⁶Note that the first stage GMM matrix \widehat{W} in (3.2) and the bootstrap distribution of J_n^{opt} shown in Figure 1, also depend on the discretization. We ignore this dependence for simplicity and compute these quantities only under the $K = 150$ grid.

⁷For instance, the estimand in (4.1) can be written in this way by setting $N(G; z) = \int h(\theta)p(z \mid \theta)dG(\theta)$ and $D(G; z) = f_G(z)$.

Anderson-Rubin-type argument, it thus suffices to test whether $L(G; c) := N(G; z) - cD(G; z) = 0$, where L is a linear functional of G . Given this reduction, the method proceeds by bias-aware inference using the affine minimax approach of Donoho (1994) and Armstrong and Kolesár (2018) carefully tailored to the empirical Bayes setting. AMARI requires a pilot F -Localization, and here we use the F -Localization implied by the constraint $J_n(f_{\tilde{G}}, \tilde{f}) \leq 13.2$ (where $\kappa = 13.2$ is the 99% quantile of the bootstrap distribution of J_n^{opt}). We refer to Ignatiadis and Wager (2022) for more details.

5.2. Fang, Santos, Shaikh, and Torgovitsky (2023) (FSST). This method was not developed for the empirical Bayes setting per say, yet here we observe that it is applicable to discrete empirical Bayes problems, such as the one here with a bivariate binomial likelihood. Using the same Anderson-Rubin-type argument as above, the confidence interval for $\theta^{\text{ratio}}(G; z)$ is the collection of values of c for which the null hypothesis that $L(G; c) = 0$ can be not rejected. Since $L(G; c)$ is a linear functional of G , after discretization as described above, this null hypothesis can be restated as

$$\exists \pi \in \mathbb{R}_+^d \text{ such that } A\pi = \beta \text{ and } a'\pi = 0 ,$$

where $d = K^2$, $\beta = f_G$, A is a $d \times p$ -dimensional matrix that encodes the bivariate binomial likelihood function, evaluated at different $z \in \mathcal{Z}$ and the grid of K^2 elements $(p_a, p_b) \in [0, 1]^2$, and $a'\pi$ encodes the restriction that $L(G; c) = 0$. Fang, Santos, Shaikh, and Torgovitsky (2023) develop a general approach to testing such a null hypothesis. See also Bai, Huang, Moon, Shaikh, and Vytlačil (2024) for related applications of this methodology in causal inference.

5.3. Further methods. There are further potential ways in which one can form confidence intervals, for instance, by pursuing the Anderson-Rubin-type argument above and test inversion. For instance, we could use CNS again (which above we used to facilitate F -Localization) to test the null hypothesis $L(G; c) = 0$, see Chernozhukov, Newey, and Santos (2023, Remark 2.3). Yet another alternative (that however, in general, lacks distribution-uniform coverage) is given in d'Haultfoeulle and Rathelot (2017).

6. ON THE CHOICE OF ESTIMAND: A COUNTERFACTUAL ODDS RATIO

Building on our previous analysis of uncertainty quantification, we now turn to a fundamental question that underlies the entire empirical Bayes approach to discrimination detection: the choice of estimand itself. The discrimination estimand in (2.3) presents two challenges that intertwine with our previous discussion of uncertainty. First, the estimand $\theta^{\text{discr}}(G; z)$ is discontinuous with respect to weak convergence of measures. This discontinuity complicates interpretation as small perturbations in the ensemble G can lead to large changes in this discrimination estimand.⁸ Second, it does not reflect the magnitude of discrimination, which, in practical policy applications, is often relevant for resource allocation and enforcement prioritization. To illustrate these concerns, consider a distribution G where $p_a = p_b + 10^{-10}$ almost surely. Then $\theta^{\text{discr}}(G; z) = 1$ for all z , suggesting complete discrimination against group b , even though such a small difference would not lead to any observable differences in hiring patterns. Moreover, if we slightly perturb G such that $p_a = p_b$ almost surely, then $\theta^{\text{discr}}(G; z) = 0$ for all z . We note that while the exact-zero discrimination threshold

⁸A similar critique also applies to common multiple testing analyses. For instance, the critique applies to the local false discovery rate $\theta^{\text{lfdr}}(G; z) := \mathbb{P}_G[\theta = 0 \mid Z = z]$ in the Gaussian empirical Bayes problem with $\theta \sim G$, $Z \mid \theta \sim N(\theta, 1)$ (McCullagh and Polson, 2018; Xiang, Ignatiadis, and McCullagh, 2024).

creates technical challenges for the estimand, this binary framing aligns with certain legal frameworks such as the Civil Rights Act, where discrimination of any magnitude is in violation of the law.

Motivated by such concerns, Kline and Walters (2021, Lemma 3) propose the logit estimand,⁹

$$\theta^{\text{logit}}(G; z) := \mathbb{E}_G [\Lambda (\Lambda^{-1}(p_a) - \Lambda^{-1}(p_b)) \mid Z = z], \quad \Lambda(p) := \frac{\exp(p)}{1 + \exp(p)}.$$

The logit estimand captures differences between group callback probabilities. Building upon this foundation, we propose a complementary counterfactual estimand that offers an alternative perspective on measuring discrimination magnitude. Our estimand answers the following question: “If we were to send additional applications to an employer with callback pattern $z = (c_a, c_b)$, what is the relative probability of observing strictly more callbacks for group a versus group b ?” Formally, for employer i with observed callback pattern $(C_a, C_b) = (c_a, c_b)$, consider the counterfactual experiment of sending L' additional applications from each group, resulting in callbacks C'_a and C'_b . Note that L' could be different from L sent in the original experiment. To define counterfactual probabilities, we assume that conditional on $\theta = (p_a, p_b)$, C'_a and C'_b are independent of C_a , C_b , and follow the bivariate binomial in (2.2) with L' trials. In this sense, we assume that our new experiment is a perfect replication of the original one except for a potentially different number of applications. See Yang, Van Zwet, Ignatiadis, and Nakagawa (2024) for a related notion of an idealized replication experiment. With this setup, we define the “posterior callback odds ratio” given $z = (c_a, c_b)$ as

$$(6.1) \quad \theta^{\text{odds}}(G; z, L') := \frac{\mathbb{P}_G[C'_a > C'_b \mid C_a = c_a, C_b = c_b]}{\mathbb{P}_G[C'_a < C'_b \mid C_a = c_a, C_b = c_b]}.$$

This estimand represents the odds ratio of callbacks for group a versus group b in a counterfactual experiment that the experimenter could actually implement. It has a natural betting interpretation: it quantifies the odds one would accept when wagering that group a will receive strictly more callbacks than group b (rather than strictly fewer) in a new experiment, given the observed callback pattern. For instance, if $\theta^{\text{odds}}(G; z, L') = 3$, a rational decision-maker would be willing to bet up to 3:1 odds on group a receiving more callbacks than group b in a counterfactual experiment with L' applications per group. This provides a meaningful quantification of discrimination that is tied to outcomes.

This estimand in (6.1) has several desirable properties, as we next document (see Appendix A for a proof).

Proposition 6.1 (Properties of posterior callback odds ratio). *Let $\theta^{\text{odds}}(G; z, L')$ be defined as in (6.1). Then:*

- (a) (No-discrimination baseline.) *If $p_a = p_b$ almost surely under G , then $\theta^{\text{odds}}(G; z, L') = 1$ for all callback patterns $z = (c_a, c_b)$.*
- (b) (Continuity under weak convergence.) *If $G_n \xrightarrow{d} G$ weakly, then $\theta(G_n; z, L') \rightarrow \theta(G; z, L')$.*
- (c) (Asymptotics with increasing number of applications L' .) *As $L' \rightarrow \infty$:*

$$\theta(G; z, L') \rightarrow \frac{\mathbb{P}_G[p_a = p_b \mid Z = z]/2 + \mathbb{P}_G[p_a > p_b \mid Z = z]}{\mathbb{P}_G[p_a = p_b \mid Z = z]/2 + \mathbb{P}_G[p_a < p_b \mid Z = z]},$$

with the convention that the right hand side is ∞ when its denominator is zero. In words, if we could send infinitely many applications in our counterfactual experiment, then the odds ratio estimand in (6.1) can be interpreted as a dampened ratio of discrimination probability $\theta^{\text{discr}}(G; z)$ in (2.3) for group a versus group b divided by the discrimination probability for group b versus group a .

⁹Kline and Walters (2021) also incorporate applicant quality in their estimand definition, which we omit.

An important property of $\theta^{\text{odds}}(G; z, L')$ is that it can be expressed as a ratio of linear functionals of G :

$$(6.2) \quad \theta^{\text{odds}}(G; (c_a, c_b), L') = \frac{\int_{[0,1]^2} \sum_{c'_b=0}^{L'-1} \sum_{c'_a=c'_b+1}^{L'} p(c'_a, c'_b \mid p_a, p_b) p(c_a, c_b \mid p_a, p_b) dG(p_a, p_b)}{\int_{[0,1]^2} \sum_{c'_b=1}^{L'} \sum_{c'_a=0}^{c'_b-1} p(c'_a, c'_b \mid p_a, p_b) p(c_a, c_b \mid p_a, p_b) dG(p_a, p_b)}.$$

Hence, this estimand is amenable to uncertainty quantification methods including F -localization described in Section 4 and other methods described in Section 5. Table 2 presents confidence intervals for this estimand using $L' = 4$ on the AGCV dataset. For the pattern $(C_a, C_b) = (1, 0)$, all three inference methods yield intervals containing 1, suggesting insufficient evidence of systematic discrimination. In contrast, for $(C_a, C_b) = (4, 0)$, all methods yield intervals strictly above 1, with AMARI providing a tighter lower bound of 17. The interpretation is operationally clear: if we send 4 more applications, we are much more likely to observe a callback pattern favoring the first applicant group.

	$(C_a, C_b) = (1, 0)$	$(C_a, C_b) = (4, 0)$
F -Localization ($\kappa = 9.9$)	0.62	8.5
AMARI	0.51	17.0
FSST	0.41	8.9

TABLE 2. Lower bounds of 95% confidence intervals for different methods for the posterior callback odds ratio estimand in (6.1) with $L' = 4$ and initial callback patterns $(C_a, C_b) = (1, 0)$, respectively, $(C_a, C_b) = (4, 0)$.

Lastly, we note that the posterior estimand is often used to inform judgments and policy decisions regarding firms with a specific callback pattern (e.g., deciding which firms to audit or sanction). For further discussion, see the work on firm discrimination in Kline, Rose, and Walters (2024), as well as related practices in teacher value-added models (Gilraine, Gu, and McMillan, 2020) and medical facility rankings (Gu and Koenker, 2023). The common empirical Bayes estimand typically takes the form of a posterior expectation, which ensures, for a suitable choice of loss function, that decisions are of high-quality on average. In the specific setting of Kline and Walters (2021), suppose $\theta^{\text{discr}}(G; z) = 0.99$, then if a policy maker were to audit all employers (or a random subset thereof) having this callback pattern z , only 1% of the resources would be spent on non-discriminating firms. However, for decisions of particularly high stakes, e.g., imposing sanctions on individual firms, the above criterion may not be sufficiently conservative and one may prefer a frequentist approach that provides error control for individual firms without relying on the exchangeability of all firms. See, e.g., the methods developed in Mogstad, Romano, Shaikh, and Wilhelm (2024) and the related discussion in Mogstad, Romano, Shaikh, and Wilhelm (2022). Even so, the empirical Bayes approach may provide useful preliminary evidence. In such use, as emphasized in our discussion above, it is additionally important to account for uncertainty from partial identification and sampling.

APPENDIX A. PROOF OF PROPOSITION 6.1

Proof.

Part (a) follows by iterated expectation and noting that C'_a is iid with C'_b conditional on any value of θ on the support of G :

$$\mathbb{P}_G[C'_a > C'_b \mid Z = z] = \mathbb{E}_G[\mathbb{P}[C'_a > C'_b \mid \theta] \mid Z = z] = \mathbb{E}_G[\mathbb{P}[C'_a < C'_b \mid \theta] \mid Z = z] = \mathbb{P}[C'_a < C'_b \mid Z = z].$$

For part (b) we may argue via representation (6.2) and proving convergence for the numerator and denominator separately. Call the numerator $N(G)$, omitting explicit dependence on z, L' and observe that $N(G) = \int \psi(p_a, p_b) dG(p_a, p_b)$, where ψ is a polynomial and thus bounded and continuous on $[0, 1]^2$. It follows that $N(G_n) \rightarrow N(G)$ when $G_n \xrightarrow{d} G$. The argument for the denominator is analogous.

For part (c), we write $\theta^{\text{odds}}(G; z, L') = \mathbb{P}_G[C'_a > C'_b, Z = z] / \mathbb{P}_G[C'_a < C'_b, Z = z]$ and again argue about the limits of the numerator and denominator separately. For the numerator, it holds that:

$$\mathbb{P}_G[C'_a > C'_b, Z = z] = \int \mathbb{P}[C'_a > C'_b, Z = z \mid \theta] dG(\theta) = \int \mathbb{P}[C'_a - C'_b > 0 \mid \theta] \mathbb{P}[Z = z \mid \theta] dG(\theta)$$

Now notice that by the central limit theorem,

$$\lim_{L' \rightarrow \infty} \mathbb{P}[C'_a - C'_b > 0 \mid \theta] = \begin{cases} 1, & \text{if } p_a > p_b, \\ 0, & \text{if } p_a < p_b, \\ 1/2, & \text{if } p_a = p_b. \end{cases}$$

By dominated convergence, it follows that

$$\mathbb{P}_G[C'_a > C'_b, Z = z] \rightarrow \int \{(\mathbb{1}(p_a = p_b)/2 + \mathbb{1}(p_a > p_b))\} \mathbb{P}[Z = z \mid \theta] dG(\theta),$$

and the right hand side the same as $\{\mathbb{P}_G[p_a = p_b \mid Z = z]/2 + \mathbb{P}_G[p_a > p_b \mid Z = z]\} \mathbb{P}_G[Z = z]$. We argue analogously regarding the denominator, and may so conclude. ■

Acknowledgments. We thank Chris Walters for helpful feedback on an earlier version of this manuscript.

REFERENCES

- ANDERSON, T. W. (1969): “Confidence Limits for the Expected Value of an Arbitrary Bounded Random Variable with a Continuous Distribution Function,” *Bulletin of the International Statistical Institute*, 43, 249–251.
- ARCEO-GOMEZ, E. O., AND R. M. CAMPOS-VAZQUEZ (2014): “Race and Marriage in the Labor Market: A Discrimination Correspondence Study in a Developing Country,” *American Economic Review*, 104(5), 376–380.
- ARMSTRONG, T. B., AND M. KOLESÁR (2018): “Optimal Inference in a Class of Regression Models,” *Econometrica*, 86(2), 655–683.
- BAI, Y., S. HUANG, S. MOON, A. M. SHAIKH, AND E. VYTLACIL (2024): “Inference for Treatment Effects Conditional on Generalized Principal Strata using Instrumental Variables,” *arXiv preprint arXiv:2411.05220*.
- 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(4), 991–1013.
- CHARNES, A., AND W. W. COOPER (1962): “Programming with Linear Fractional Functionals,” *Naval Research Logistics Quarterly*, 9(3-4), 181–186.
- CHERNOZHUKOV, V., W. K. NEWEY, AND A. SANTOS (2023): “Constrained Conditional Moment Restriction Models,” *Econometrica*, 91(2), 709–736.
- COEY, D., AND K. HUNG (2022): “Empirical Bayes Selection for Value Maximization,” *arXiv preprint arXiv:2210.03905*.

- D’HAULTFOEUILLE, X., AND R. RATHELOT (2017): “Measuring segregation on small units: A partial identification analysis,” *Quantitative Economics*, 8(1), 39–73.
- DONOHO, D. L. (1994): “Statistical Estimation and Optimal Recovery,” *The Annals of Statistics*, pp. 238–270.
- EFRON, B. (2019): “Bayes, Oracle Bayes and Empirical Bayes,” *Statistical Science*, 34(2), 177–201.
- FANG, Z., A. SANTOS, A. M. SHAIKH, AND A. TORGOVITSKY (2023): “Inference for Large-Scale Linear Systems With Known Coefficients,” *Econometrica*, 91(1), 299–327.
- GILRAINE, M., J. GU, AND R. McMILLAN (2020): “A new method for estimating teacher value-added,” Discussion paper, National Bureau of Economic Research.
- GU, J., AND R. KOENKER (2022): “Ranking and selection from pairwise comparisons: empirical Bayes methods for citation analysis,” in *AEA Papers and Proceedings*, vol. 112, pp. 624–629. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.
- (2023): “Invidious comparisons: Ranking and selection as compound decisions,” *Econometrica*, 91(1), 1–41.
- IGNATIADIS, N., AND S. WAGER (2022): “Confidence intervals for nonparametric empirical Bayes analysis,” *Journal of the American Statistical Association*, 117(539), 1149–1166.
- IMBENS, G. (2022): “Comment on: “Confidence Intervals for Nonparametric Empirical Bayes Analysis” by Ignatiadis and Wager,” *Journal of the American Statistical Association*, 117(539), 1181–1182.
- KLINE, P., AND C. WALTERS (2021): “Reasonable Doubt: Experimental Detection of Job-level Employment Discrimination,” *Econometrica*, 89(2), 765–792.
- KLINE, P. M., E. K. ROSE, AND C. R. WALTERS (2024): “A discrimination report card,” *American Economic Review*, 114(8), 2472–2525.
- KOENKER, R., AND I. MIZERA (2014): “Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules,” *Journal of the American Statistical Association*, 109(506), 674–685.
- LORD, F. M., AND N. CRESSIE (1975): “An empirical Bayes procedure for finding an interval estimate,” *Sankhyā: The Indian Journal of Statistics, Series B*, pp. 1–9.
- LORD, F. M., AND M. L. STOCKING (1976): “An Interval Estimate for Making Statistical Inferences about True Scores,” *Psychometrika*, 41(1), 79–87.
- MASSART, P. (1990): “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality,” *The Annals of Probability*, pp. 1269–1283.
- MCCULLAGH, P., AND N. G. POLSON (2018): “Statistical Sparsity,” *Biometrika*, 105(4), 797–814.
- METCALFE, R. D., A. B. SOLLACI, AND C. SYVERSON (2023): “Managers and Productivity in Retail,” Working Paper 31192, National Bureau of Economic Research.
- MOGSTAD, M., J. ROMANO, A. SHAIKH, AND D. WILHELM (2022): “Comment on” Invidious Comparisons: Ranking and Selection as Compound Decisions,” *Econometrica*.
- MOGSTAD, M., J. P. ROMANO, A. M. SHAIKH, AND D. WILHELM (2024): “Inference for ranks with applications to mobility across neighbourhoods and academic achievement across countries,” *Review of Economic Studies*, 91(1), 476–518.
- 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,” *The BE Journal of Economic Analysis & Policy*, 15(3), 1093–1125.

- ROBBINS, H. (1956): “An Empirical Bayes Approach to Statistics,” in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 157–163. The Regents of the University of California.
- ROMANO, J. P., AND A. M. SHAIKH (2008): “Inference for identifiable parameters in partially identified econometric models,” *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.
- ROZEMA, K., AND M. SCHANZENBACH (2019): “Good cop, bad cop: Using civilian allegations to predict police misconduct,” *American Economic Journal: Economic Policy*, 11(2), 225–268.
- SCHEFFÉ, H. (1953): “A Method for Judging All Contrasts in the Analysis of Variance,” *Biometrika*, 40(1-2), 87–110.
- STOYE, J. (2009): “More on Confidence Intervals for Partially Identified Parameters,” *Econometrica*, 77(4), 1299–1315.
- WALTERS, C. (2024): “Empirical Bayes methods in labor economics,” in *Handbook of Labor Economics*, vol. 5, pp. 183–260. Elsevier.
- WERNERFELT, N., A. TUCHMAN, B. SHAPIRO, AND R. MOAKLER (2022): “Estimating the value of offsite data to advertisers on meta,” *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (114).
- WOOD, G. R. (1999): “Binomial Mixtures: Geometric Estimation of the Mixing Distribution,” *The Annals of Statistics*, 27(5), 1706–1721.
- XIANG, D., N. IGNATIADIS, AND P. MCCULLAGH (2024): “Interpretation of Local False Discovery Rates under the Zero Assumption,” *arXiv preprint*, arXiv:2402.08792.
- YANG, Y., E. VAN ZWET, N. IGNATIADIS, AND S. NAKAGAWA (2024): “A Large-Scale in Silico Replication of Ecological and Evolutionary Studies,” *Nature Ecology & Evolution*, 8(12), 2179–2183.

DEPARTMENT OF ECONOMICS, UNIVERSITY OF TORONTO

Email address: `jiaying.gu@utoronto.ca`

DEPARTMENT OF STATISTICS AND DATA SCIENCE INSTITUTE, UNIVERSITY OF CHICAGO

Email address: `ignat@uchicago.edu`

DEPARTMENT OF ECONOMICS, UNIVERSITY OF CHICAGO

Email address: `amshaikh@uchicago.edu`