Fast Threshold Tests for Detecting Discrimination

Emma Pierson [1] **Sam Corbett-Davies** [1] **Sharad Goel** [1]

Abstract

Threshold tests have recently been proposed as a robust method for detecting bias in lending, hiring, and policing decisions. For example, in the case of credit extensions, these tests aim to estimate the bar for granting loans to white and minority applicants, with a higher inferred threshold for minorities indicative of discrimination. This technique, however, requires fitting a Bayesian latent variable model for which inference is often computationally challenging. Here we develop a method for fitting threshold tests that is more than 75 times faster than the existing approach, reducing computation from hours to minutes. We demonstrate this technique by analyzing 2.7 million police stops of pedestrians in New York City between 2008 and 2012. To achieve these performance gains, we introduce and analyze a flexible family of probability distributions on the interval $[0, 1]$—which we call *discriminant distributions*—that is computationally efficient to work with. These discriminant distributions may aid inference in a variety of applications beyond threshold tests.

1. Introduction

There is wide interest in detecting and quantifying bias in human decisions, but well-known problems with traditional statistical tests of discrimination have hampered rigorous analysis. The primary goal of such work is to determine whether decision-makers apply different standards to groups defined by race, gender, or other protected attributes—what economists call *taste-based discrimination* (Becker, 1957). For example, in the context of banking, such discrimination might mean that minorities are granted loans only when they are exceptionally creditworthy. The key statistical challenge is that an individual's qualifications are typically only partially observed (e.g., researchers may not know an applicant's full credit history); it is thus unclear whether observed disparities are attributable to discrimination or omitted variables.

To address this problem, Simoiu et al. (2017) recently proposed the *threshold test*, which considers both the decisions made (e.g., whether a loan was granted) and the outcomes of those decisions (e.g., whether a loan was repaid). With this information, the test simultaneously estimates decision thresholds and risk profiles via a Bayesian latent variable model. This approach is theoretically appealing and mitigates some of the most serious statistical shortcomings of past methods. Fitting the model, however, is computationally challenging, often requiring several hours on moderately sized datasets. As is common in fully Bayesian inference, the threshold model is typically fit with Hamiltonian Monte Carlo (HMC) sampling. In this case, HMC involves repeatedly evaluating gradients of conditional beta distributions that are expensive to compute.

Here we introduce and analyze a family of distributions on the interval $[0, 1]$—which we call *discriminant distributions*—that is efficient for carrying out common statistical operations. By replacing the beta distributions in the threshold test with discriminant distributions, we can speed up inference by two orders of magnitude. Discriminant distributions comprise a natural subset of logit-normal mixture distributions, and this family is sufficiently expressive to approximate logit-normal and beta distributions for a wide range of parameters.

To demonstrate our method, we analyze 2.7 million police stops of pedestrians in New York City between 2008 and 2012. We specifically apply the threshold test to assess possible bias in decisions to search individuals for weapons. We further modify the threshold test to detect discrimination in the decision to stop an individual. In this case—and in contrast to search decisions—one does not observe those who were not stopped. The setup thus bears resemblance to learning classifiers from only positive examples (Elkan & Noto, 2008; Mordelet & Vert, 2014; du Plessis et al., 2014). On both problems (search decisions and stop decisions), our method accelerates inference by more than 75-fold.

2. Background

2.1. Traditional statistical tests of discrimination

To motivate the threshold test, we briefly discuss two traditional statistical tests of discrimination: the *benchmark*

[1]Stanford University, Stanford, CA 94305

test (or *benchmarking*) and the *outcome test*. The benchmark test analyzes the rate at which some action is taken (e.g., the rate at which stopped pedestrians are searched). Decision rates might vary across racial groups for a variety of legitimate reasons, such as race-specific differences in criminal behavior. One thus attempts to estimate decision rates after controlling for all legitimate factors, typically via a regression model. If decision rates still differ by race after such conditioning, the benchmark test would suggest bias. Though popular, this test suffers from the well-known problem of *omitted variable bias*, as it is typically impossible for researchers to observe—and hence control for—all legitimate factors that might affect decisions. For example, if evasiveness is a reliable indicator of contraband, is not observed by researchers, and is differentially distributed across race groups, the benchmark test might indicate discrimination where there is none. This concern is especially problematic for face-to-face interactions such as police stops that may rely on hard-to-quantify behavioral indicators.

Addressing this shortcoming of benchmarking, Becker (1993; 1957) proposed the outcome test, which is based not on the rate at which decisions are made but on the *hit rate* (i.e, the success rate) of those decisions. Becker reasoned that even if one cannot observe the exact rationale for a search, absent discrimination weapons should be found on searched minorities at the same rate as on searched whites. If searches of minorities turn up weapons at lower rates than searches of whites, it suggests that officers are applying a double standard, searching minorities on the basis of less evidence.

Outcome tests, however, are also imperfect measures of discrimination (Ayres, 2002). Suppose that there are two, easily distinguishable types of white pedestrians: those who have a 1% chance of carrying weapons, and those who have a 75% chance. Similarly assume that black pedestrians have either a 1% or 50% chance of carrying weapons. If officers, in a race-neutral manner, search individuals who are at least 10% likely to be carrying a weapon, then searches of whites will be successful 75% of the time whereas searches of blacks will be successful only 50% of the time. With such a race-neutral threshold, no individual is treated differently because of their race. Thus, contrary to the findings of the outcome test (which identifies discrimination against blacks due to their lower hit rate), no discrimination is present. This simple example illustrates a subtle failure of outcome tests known as the problem of infra-marginality (Ayres, 2002; Simoiu et al., 2017; Anwar & Fang, 2006; Engel, 2008)

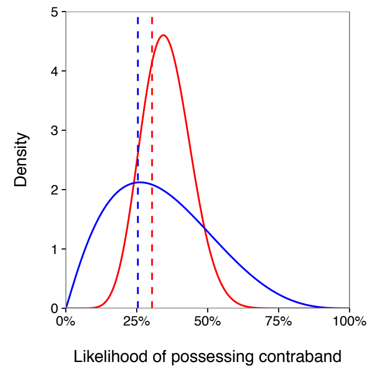

Figure 1. An illustration of hypothetical risk distributions (solid curves) and search thresholds (dashed vertical lines). In this example, the blue group is searched at a lower threshold than the red group, and so the blue group by definition faces discrimination.

2.2. The threshold test

To circumvent this problem of infra-marginality, the *threshold test* of Simoiu et al. (2017) attempts to directly infer race-specific search thresholds. Though still relatively new, the test has already been used to analyze tens of millions of police stops across the United States (Pierson et al., 2017). The threshold test is based on a Bayesian latent variable model that formalizes the following stylized process of search and discovery. Upon stopping a pedestrian, officers observe the probability p the individual is carrying a weapon; this probability summarizes all the available information, such as the stopped individual's age and gender, his criminal record, and behavioral indicators like nervousness and evasiveness. Because these probabilities vary from one individual to the next, p is modeled as being drawn from a *risk distribution* that depends on the stopped person's race (r) and the location of the stop (d), where location might indicate the precinct in which the stop occurred. Officers deterministically conduct a search if the probability p exceeds a race- and location-specific threshold (t_{rd}), and if a search is conducted, a weapon is found with probability p.

By reasoning in terms of risk distributions, one avoids the omitted variables problem by marginalizing out all unobserved variables. In this formulation, one need not observe the factors that led to any given decision, and can instead infer the aggregate distribution of risk for each group.

Figure 1 illustrates hypothetical risk distributions and thresholds for two groups in a single location. This representation allows us to visually describe the mapping from thresholds and risk distributions to search rates and hit rates. Suppose P_{rd} is a random variable that gives the probability of finding a weapon on a stopped pedestrian in group r in location d. The search rate s_{rd} of group r in location d is $\Pr(P_{rd} > t_{rd})$, the probability a randomly selected pedestrian in that group and location exceeds the race-specific search threshold; graphically, this is the proportion of the risk distribution to the right of the threshold. The hit rate is the probability that a random *searched* pedestrian is carrying a weapon: $h_{rd} = \mathbb{E}[P_{rd} \mid P_{rd} > t_{rd}]$; in other words, the hit rate is the mean of the risk distribution conditional on being above the threshold.

The primary goal of inference is to determine the latent thresholds t_{rd}. If the thresholds applied to one race group are consistently lower than the thresholds applied to another, this suggests discrimination against the group with the lower thresholds.

In order to estimate the decision thresholds, the risk distributions must be simultaneously inferred. In Simoiu et al. (2017), these risk profiles take the form of beta distributions parameterized by means $\phi_{rd} = \text{logit}^{-1}(\phi_r + \phi_d)$ and total count parameters $\lambda_{rd} = \exp(\lambda_r + \lambda_d)$, where ϕ_r, ϕ_d, λ_r, and λ_d are parameters that depend on the race of the stopped individuals and the location of the stops.[2] Reparameterizing these risk distributions is the key to accelerating inference.

Given the number of searches and hits by race and location, we can compute the likelihood of the observed data under any set of model parameters $\{\phi_d, \phi_r, \lambda_d, \lambda_r, t_{rd}\}$. In the Bayesian framework of the threshold test, one can likewise compute the posterior distribution of the parameters given the data and prior distributions.

2.3. Inference via Hamiltonian Monte Carlo

Bayesian inference is challenging when the parameter space is high dimensional, because random walk MCMC methods fail to fully explore the complex posterior in any reasonable time. This obstacle can be addressed with Hamiltonian Monte Carlo (HMC) methods (Neal et al., 2011; Betancourt & Girolami, 2015; Chen et al., 2014), which propose new samples by numerically integrating along the gradient of the log posterior, allowing for more efficient exploration. The speed of convergence of HMC depends on three factors: (1) the gradient computation time per integration step; (2) the number of integration steps per sample; and (3) the number of effectively independent samples relative to the total number of samples (i.e., the effective sample size). The first can be improved by simplifying the analytical form of the log posterior and its derivatives. The second and third factors depend on the geometry of the posterior: a smooth posterior allows for longer paths between samples that take fewer integration steps to traverse (Betancourt, 2017). On all three measures, our new threshold model generally outperforms that of Simoiu et al. (2017); the improvement in gradient computation is particularly substantial.

[2] In terms of the standard count parameters α and β of the beta distribution, $\phi = \alpha/(\alpha + \beta)$ and $\lambda = \alpha + \beta$.

3. Discriminant distributions

The computational complexity of the standard threshold test is in large part due to difficulties of working with beta distributions. Specifically, when P has a beta distribution, it is expensive to compute the search rate $\Pr(P > t)$, the hit rate $\mathbb{E}[P \mid P > t]$, and their associated derivatives (Boik & Robison-Cox, 1998). Here we introduce an alternative family of *discriminant distributions* for which it is efficient to compute these quantities. We motivate and analyze this family in the specific context of the threshold test, but the family itself can be applied much more widely.

To define discriminant distributions, we first imagine that there are two classes (positive and negative), and assume the probability of being in the positive class is ϕ. For example, positive examples might correspond to individuals who are carrying weapons, and negative examples to those who are not. We further assume that each class emits *signals* that are normally distributed according to $\text{N}(\mu_0, \sigma_0)$ and $\text{N}(\mu_1, \sigma_1)$, respectively. Denote by X the signal emitted by a random instance in the population, and by $Y \in \{0, 1\}$ its class membership. Then, given an observed signal x, one can compute the probability $g(x) = \Pr(Y = 1 \mid X = x)$ that it was emitted by a member of the positive class. Throughout the paper, we will term the domain of g the *signal space* and its range the *probability space*. Finally, we say the random variable $g(X)$ has a discriminant distribution with parameters ϕ, μ_0, σ_0, μ_1, and σ_1. Definition 3.1 formalizes this characterization of discriminant distributions.

Definition 3.1 (Discriminant distribution)**.** Consider parameters $\phi \in (0, 1)$, $\mu_0 \in \mathbb{R}$, $\sigma_0 \in \mathbb{R}_+$, $\mu_1 \in \mathbb{R}$, and $\sigma_1 \in \mathbb{R}_+$, where $\mu_1 > \mu_0$. Then the discriminant distribution $\text{disc}(\phi, \mu_0, \sigma_0, \mu_1, \sigma_1)$ is defined as follows. Suppose

$$Y \sim \text{Bernoulli}(\phi),$$

and

$$X \mid Y = 0 \sim \text{N}(\mu_0, \sigma_0)$$
$$X \mid Y = 1 \sim \text{N}(\mu_1, \sigma_1).$$

Set $g(x) = \Pr(Y = 1 \mid X = x)$. Then the random variable $g(X)$ is distributed as $\text{disc}(\phi, \mu_0, \sigma_0, \mu_1, \sigma_1)$.

Our description above mirrors the motivation of linear discriminant analysis (LDA). Although it is common to consider the conditional probability of class membership $g(x)$, it appears less common to consider the *distribution* of these probabilities as an alternative to beta or logit-normal distributions. Moreover, to the best of our knowledge, the computational properties of discriminant distributions have not been previously studied.

As in the case of LDA, the statistical properties of discriminant distributions are particularly nice when the underlying normal distributions have the same variance. Proposition 3.1 below establishes a key monotonicity property that is standard in the development of LDA; though the statement and derivation are well-known in that setting, we include them here for completeness.

Proposition 3.1 (Monotonicity)**.** Given a discriminant distribution $\mathrm{disc}(\phi, \mu_1, \sigma_1, \mu_2, \sigma_2)$, the mapping g from signal space to probability space is monotonic if and only if $\sigma_0 = \sigma_1$.

Proof. The mapping from signal space to probability space can be found with Bayes' rule:

$$\begin{aligned} g(x) &= \Pr(Y = 1 \mid X = x) \\ &= \frac{\Pr(Y=1)\mathrm{N}(x;\mu_1,\sigma_1)}{\Pr(Y=1)\mathrm{N}(x;\mu_1,\sigma_1) + \Pr(Y=0)\mathrm{N}(x;\mu_0,\sigma_0)} \\ &= \mathrm{logit}^{-1}(Ax^2 + Bx + C) \end{aligned}$$

where

$$A = \frac{\sigma_1^2 - \sigma_0^2}{2\sigma_0^2\sigma_1^2} \qquad B = \frac{\mu_1\sigma_0^2 - \mu_0\sigma_1^2}{\sigma_0^2\sigma_1^2}$$
$$C = \frac{\sigma_1^2\mu_0^2 - \sigma_0^2\mu_1^2}{2\sigma_0^2\sigma_1^2} - \log\left(\frac{1-\phi}{\phi}\frac{\sigma_1}{\sigma_0}\right).$$

This is the composition of a quadratic in x and the inverse logit function, which will be monotonic if and only if the quadratic is monotonic, requiring $\sigma_0^2 - \sigma_1^2 = 0$. □

Homoskedastic discriminant distributions (i.e., with $\sigma_0 = \sigma_1$) by definition involve four parameters. The following proposition shows that in fact only two parameters are required to fully describe this family of distributions. This simplified parameterization is particularly useful for computation.

Proposition 3.2 (2-parameter representation)**.** Suppose $\mathrm{disc}(\phi, \mu_1, \sigma, \mu_2, \sigma)$ and $\mathrm{disc}(\phi', \mu_1', \sigma', \mu_2', \sigma')$ are two homoskedastic discriminant distributions. Let

$$\delta = \frac{\mu_1 - \mu_0}{\sigma}$$

and define δ' analogously. Then the two distributions are identical if $\phi = \phi'$ and $\delta = \delta'$. As a result, homoskedastic discriminant distributions can be parameterized by ϕ and δ alone.

Proof. We establish this result by explicitly deriving the density of $g(X)$, where g is the usual mapping from signal to probability space. Doing so first requires computing the inverse transformation from probability space to signal space:

$$g^{-1}(p) = \frac{\mu_1^2 - \mu_0^2 - 2\sigma^2\log\left(\frac{\phi}{1-\phi}\frac{1-p}{p}\right)}{2(\mu_1 - \mu_0)}.$$

Now,

$$\frac{d}{dp}(g^{-1}(p)) = \frac{\sigma^2}{p(1-p)(\mu_1 - \mu_0)}.$$

The density of X is

$$f_X(x) = \phi N(x;\mu_1,\sigma_1) + (1-\phi)N(x;\mu_0,\sigma_0).$$

We can accordingly compute the density of $g(X)$ by the change of variables formula:

$$\begin{aligned} f_P(p) &= f_X(g^{-1}(p))\left|\frac{d}{dp}(g^{-1}(p))\right| \\ &= \frac{\sigma}{\sqrt{2\pi}p(1-p)(\mu_1-\mu_0)}\left[\phi\exp\left(-\frac{\left(\alpha + \frac{(\mu_1-\mu_0)^2}{\sigma^2}\right)^2}{8\frac{(\mu_1-\mu_0)^2}{\sigma^2}}\right)\right. \\ &\qquad \left. + (1-\phi)\exp\left(-\frac{\left(\alpha - \frac{(\mu_1-\mu_0)^2}{\sigma^2}\right)^2}{8\frac{(\mu_1-\mu_0)^2}{\sigma^2}}\right)\right], \end{aligned}$$

where

$$\alpha = 2\log\left(\frac{\phi}{1-\phi}\frac{1-p}{p}\right).$$

Without loss of generality we can set $\delta = (\mu_1 - \mu_0)/\sigma$, resulting in a 2-parameter family of densities:

$$\begin{aligned} f_P(p) = \frac{1}{\sqrt{2\pi}p(1-p)\delta}&\left[\phi\exp\left(-\frac{\left(\alpha+\delta^2\right)^2}{8\delta^2}\right)\right. \\ &\left. + (1-\phi)\exp\left(-\frac{\left(\alpha-\delta^2\right)^2}{8\delta^2}\right)\right]. \end{aligned} \tag{1}$$

□

Given Proposition 3.2, we henceforth write $\mathrm{disc}(\phi, \delta)$ to denote a homoskedastic discriminant distribution. Even though the distribution itself depends on only these two parameters, the transformation g from signal space to probability space depends on the particular 4-parameter representation we use. For simplicity, we consider the representation with $\mu_0 = 0$ and $\sigma = 1$. This choice yields the simplified transformation function:

$$g(x) = \frac{1}{1 + \frac{1-\phi}{\phi}\exp\left(-\delta x + \delta^2/2\right)}.$$

Our primary motivation for introducing discriminant distributions was to accelerate key computations of the (complementary) CDF and conditional means. Letting $P = g(X)$, we are specifically interested in computing $\Pr(P > t)$ and

$\mathbb{E}[P \mid P > t]$. With (homoskedastic) discriminant distributions, these quantities map nicely to signal space, where they can be computed efficiently.

Denote by $\Phi(x; \mu, \sigma)$ the normal CDF. Then the complementary CDF of P can be computed as follows.

$$\begin{aligned}
\Pr(P > t) &= \Pr\left(X > g^{-1}(t)\right) \\
&= \Pr(Y = 0, X > g^{-1}(t)) + \\
&\quad \Pr(Y = 1, X > g^{-1}(t)) \\
&= (1 - \phi)\Phi(g^{-1}(t); 0, 1) + \phi\Phi(g^{-1}(t); \delta; 1).
\end{aligned}$$

For the conditional mean, we have

$$\begin{aligned}
\mathbb{E}[P \mid P > t] &= \Pr\left(Y = 1 \mid X > g^{-1}(t)\right) \\
&= \frac{\Pr(Y = 1, X > g^{-1}(t))}{\Pr(X > g^{-1}(t))} \\
&= \frac{\phi\Phi(g^{-1}(t); \delta; 1)}{(1 - \phi)\Phi(g^{-1}(t); 0, 1) + \phi\Phi(g^{-1}(t); \delta; 1)}.
\end{aligned}$$

Importantly, the CDF and conditional means for discriminant distributions are closely related to those for the normal distributions, and as such are computationally efficient to work with. In particular, the gradients of these functions are relatively straightforward to evaluate. In contrast, the corresponding quantities for logit-normal and beta distributions involve tricky numerical approximations (Frederic & Lad, 2008; Jones, 2009).

Finally, we show that discriminant distributions are an expressive family of distributions. Figure 2 shows that discriminant distributions can approximate typical instantiations of the logit-normal and beta distributions. First, we select the parameters of the reference distribution: (μ, σ) for the logit-normal or (ϕ, λ) for the beta. Then we numerically optimize the parameters of the discriminant distribution to minimize the total variation distance between the reference distribution and the discriminant distribution. The top row of Fig. 2 shows some typical densities and their approximations. The bottom row systematically investigates the approximation error for a wide range of parameter values. The discriminant distribution fits the logit-normal very well (distance below 0.1 for all distributions with $\sigma \leq 3$), and approximates the beta distribution moderately well (distance below 0.2 for $\lambda \geq 1$).

Discriminant distributions approximate logit-normal distributions particularly well because they in fact form a subset of logit-normal mixture distributions. To see this, consider the following rearrangement of the first term in the mixture

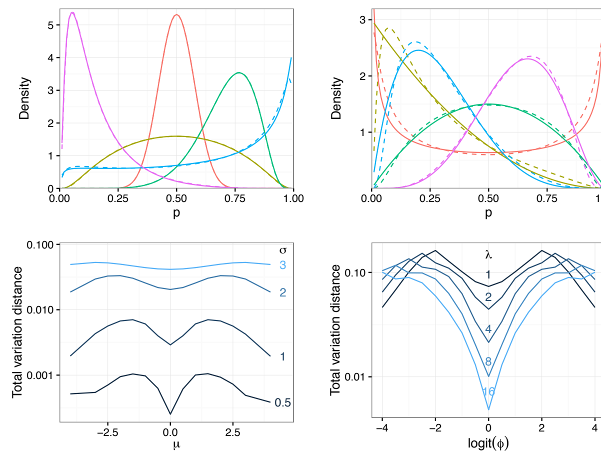

Figure 2. The discriminant distribution can approximate the $\mathrm{logit\text{-}normal}(\mu, \sigma)$ (left panels) and $\mathrm{beta}(\phi, \lambda)$ (right panels) distributions. This can be shown by approximating various instantiations of the beta and logit-normal distributions with the best fit discriminant distribution, found by minimizing the total variation distance between the distributions. In the top row we illustrate some typical distributions (solid lines) and their discriminant distribution approximations (dashed lines, some approximations are too accurate for these lines to be seen). In the bottom row we explore a wide range of parameter values for distributions and report the total variation distance between the reference distribution and its discriminant distribution approximation.

in Eq. (1):

$$\begin{aligned}
&\frac{\exp\left(-\frac{\left(2\log\left(\frac{\phi}{1-\phi}\frac{1-p}{p}\right)+\delta^2\right)^2}{8\delta^2}\right)}{\sqrt{2\pi}p(1-p)\delta} \\
&\quad = \frac{\exp\left(-\frac{\left(\mathrm{logit}(p)-\left(\mathrm{logit}(\phi)+\frac{\delta^2}{2}\right)\right)^2}{2\delta^2}\right)}{\sqrt{2\pi}p(1-p)\delta}.
\end{aligned}$$

This is the density of a logit-normal with parameters $\mu = \mathrm{logit}(\phi) + \frac{\delta^2}{2}$ and $\sigma = \delta$. Thus, we can express the discriminant distribution as a specific 2-parameter mixture of logit-normals (where $f_l(x; \mu, \sigma)$ is the density function of the logit-normal):

$$\begin{aligned}
f_P(p) = \phi f_l &\left(p; \mathrm{logit}(\phi) + \frac{\delta^2}{2}, \delta\right) \\
&+ (1 - \phi) f_l \left(p; \mathrm{logit}(\phi) - \frac{\delta^2}{2}, \delta\right).
\end{aligned}$$

In this form, we can see that for small δ or ϕ close to 0 or 1, the distribution is almost equivalent to a (single) logit-normal.

4. Testing for discrimination in New York City's stop-and-frisk policy

To demonstrate the value of discriminant distributions for speeding up the threshold test, we analyze a dataset of pedestrian stops conducted by New York City police officers under its "stop-and-frisk" practice. Officers have a legal right to stop and briefly detain individuals when they suspect criminal activity. There is worry, however, that such discretionary decisions are prone to racial bias; indeed the NYPD practice was recently ruled discriminatory in federal court and subsequently curtailed (*Floyd v. City of New York*, 2013). Here we revisit the statistical evidence for discrimination.

Our dataset contains information on 2.7 million police stops occurring between 2008 and 2012. Several variables are available for each stop, including the race of the pedestrian, the police precinct in which the stop occurred, whether the pedestrian was "frisked" (i.e., patted-down in search of a weapon), and whether a weapon was found. We confine our analysis to stops of white, black, and Hispanic pedestrians, as there are relatively few stops involving individuals of other races.

We use the threshold test to analyze two specific decisions: the initial stop decision, and the subsequent decision of whether or not to conduct a frisk. Analyzing frisk decisions is a straightforward application of the threshold test. In this case, we show that simply replacing beta distributions in the model with discriminant distributions results in more than a 100-fold speedup. Analyzing stop decisions requires extending the threshold model to the case where one does not observe negative examples (i.e., one does not observe those who were not stopped). After describing this extension, we show that discriminant distributions again produce significant speedups.

4.1. Testing for discrimination in frisk decisions

When using beta distributions in the threshold test, it takes nearly two hours to infer the model parameters; when we replace beta distributions with discriminant distributions, inference completes in under one minute.[3] The inferred thresholds under both models are nearly identical (correlation = 0.95), and indicate discrimination against black and Hispanic individuals.

We plot the inferred thresholds (under the accelerated model) in Figure 3. Each point corresponds to the threshold for one precinct; we plot the threshold for white pedestrians on the horizontal axis, and for minority pedestrians on the vertical axis. Thresholds for minority pedestrians are consistently lower than thresholds for white pedestrians within the same precinct, suggestive of discrimination.

[3]We fit the threshold models using Stan (Carpenter et al., 2016), an open-source modeling language for full Bayesian statistical inference via HMC.

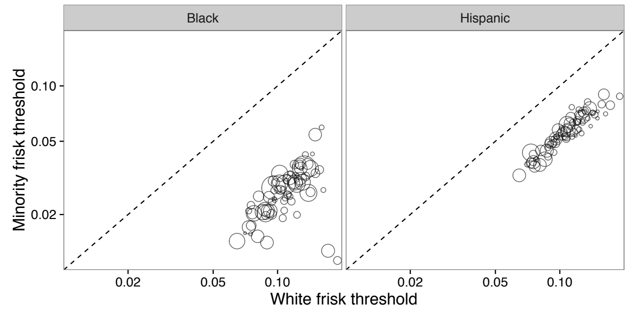

Figure 3. Inferred thresholds for frisks in the stop-and-frisk data. The thresholds for white pedestrians are plotted on the horizontal axis and the thresholds for minority pedestrians in the same precinct are plotted on the vertical axis. The dotted line denotes equal thresholds. The size of each circle corresponds to the number of stops of minority pedestrians. Axes are logarithmic.

Such evidence of discrimination is consistent with results from past statistical analyses of New York City's stop-and-frisk practices (Goel et al., 2016; Gelman et al., 2007). The threshold test, however, mitigates concern that previous findings were driven by infra-marginality. We note that Simoiu et al. (2017) find infra-marginality likely caused traditional tests of discrimination to reach spurious conclusions, making it important to explicitly account for this possibility.

As noted above, the accelerated threshold test produces parameter estimates that are nearly identical to those under the traditional test. We further confirm that these inferred parameters capture key features of the data by performing posterior predictive checks (Gelman et al., 1996). Specifically, we compute the model-inferred frisk and hit rates for each precinct and race group, and compare these to the observed rates (Figure 4). The model almost perfectly fits the observed frisk rates and fits the observed hit rates quite well; both statistics also exhibit a lack of systematic bias.

Why is it that discriminant distributions result in such a dramatic increase in performance? The time to compute a given number of effective samples is the product of three terms:

$$\frac{\text{seconds}}{n_{\text{eff}}} = \frac{\text{samples}}{n_{\text{eff}}} \cdot \frac{\text{integration steps}}{\text{sample}} \cdot \frac{\text{seconds}}{\text{integration steps}}.$$

As shown in Table 1, we find that all three factors are significantly reduced by using discriminant distributions, with the final term providing the most significant reduction. In particular, using discriminant distributions reduces the time per effective sample by a factor of 760. In practice, one typically runs chains in parallel, and so total running time is determined by the last chain to terminate. When running

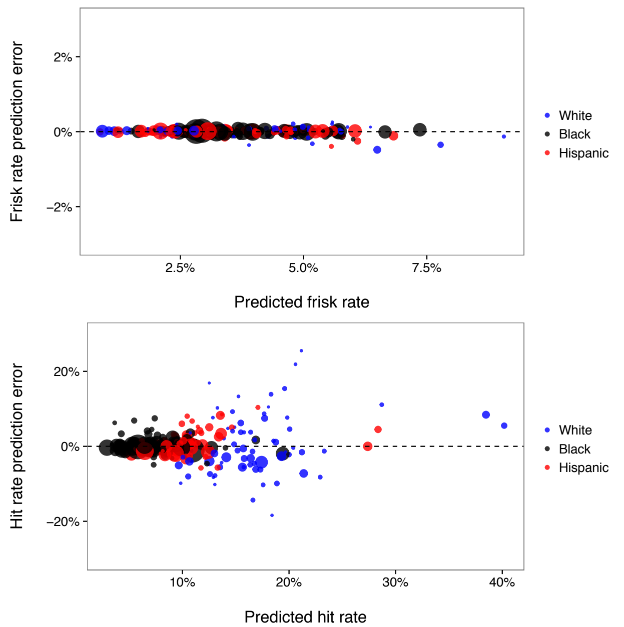

Figure 4. Posterior predictive checks for frisk rate and hit rate in stop-and-frisk data. Each point represents one precinct-race pair. Points are sized by the number of stops for that precinct and race.

five chains in parallel for 5,000 iterations each, the accelerated model is faster by a factor of 140.

4.2. Testing for discrimination in stop decisions

We now extend the threshold model to test for discrimination in an officer's decision to stop a pedestrian. In contrast to frisk decisions, we do not observe instances in which an officer decided *not* to carry out a stop. Inferring thresholds with such censored data is analogous to learning classifiers from only positive and unlabeled examples (Elkan & Noto, 2008; Mordelet & Vert, 2014; du Plessis et al., 2014). To overcome this problem, we assume officers are equally likely to encounter anyone in a precinct. Coupled with demographic data compiled by the U.S. Census Bureau, this assumption lets us estimate the racial distribution of individuals encountered by officers. Such estimates are imperfect proxies, in part because residential populations differ from daytime populations (Bhaduri, 2008); we find, however, that our inferences are robust to reasonable violations of this assumption.

4.2.1. Model

When analyzing frisk decisions, the decision itself and the success of a frisk were modeled as random outcomes. In the case of stops, we model as random the race of stopped individuals and whether a stop was successful (i.e., turned up a weapon). The likelihood of the observed outcomes can then be computed under any set of model parameters, which in turn allows us to compute posterior parameter estimates.

Table 1. Breakdown of sampling times for the frisk decision model. Each row reports the improvement for the discriminant distribution model relative to the beta distribution model followed by the statistics for both models. We report the seconds per effective sample (first row), which is the product of the numbers in the second to fourth rows. We also report the time to fit the models used in our analysis (final row) using 5 chains run in parallel for 5,000 samples.

	Speedup	Beta	Disc
Seconds/n_{EFF}	**760x**	26	0.3
Samples/n_{EFF}	**5x**	22	5
Integration steps/sample	**3x**	152	49
Seconds/integration step	**53x**	.008	.0001
Seconds to fit model	**143x**	5977	42

Let S_{rd} denote the number of stops of individuals of race r in precinct d, and let H_{rd} denote the number of such stops that yield a weapon. We denote by c_{rd} the fraction of people in a precinct of a given race. Letting R_d denote the race of an individual randomly encountered by the police in precinct d, we have

$$\Pr(R_d = r \mid \text{stopped}) \propto \Pr(\text{stopped} \mid R_d = r) \Pr(R_d = r).$$

Assuming officers are equally likely to encounter everyone in a precinct, $\Pr(R_d = r) \propto c_{rd}$. We further assume that individuals of race r are stopped when their probability of carrying a weapon exceeds a race- and precinct-specific threshold t_{rd}; this assumption is analogous to the one made for the frisk model. Setting $\theta_{rd} = \Pr(R_d = r \mid \text{stopped})$, we have

$$\theta_{rd} \propto c_{rd} \cdot \Pr(\text{stopped} \mid R_d = r)$$

where

$$\begin{aligned}\Pr(\text{stopped} \mid R_d = r) &= \Pr(P_{rd} > t_{rd}) \\ &= (1 - \phi_{rd})(1 - \Phi(t_{rd}; 0, 1)) \\ &\quad + \phi_{rd}(1 - \Phi(t_{rd}; 0, \delta_{rd})).\end{aligned}$$

For each precinct d, the racial composition of stops is thus distributed as a multinomial:

$$\vec{S}_d \sim \text{multinomial}\left(\vec{\theta}_d, N_d\right)$$

where N_d denotes the total number of stops conducted in that precinct, $\vec{\theta}_d$ is a vector of race-specific stop probabilities θ_{rd}, and $\vec{S}_d$ is the number of stops of each race group in that precinct. We model hits as in the frisk model. To complete the full Bayesian specification, we put normal or half-normal priors on all the parameters: $\{\phi_d, \phi_r, \lambda_d, \lambda_r, t_{rd}\}$.[4] Figure 5 presents a graphical representation of the generative process.

[4]Following Simoiu et al. (2017) and Pierson et al. (2017), we

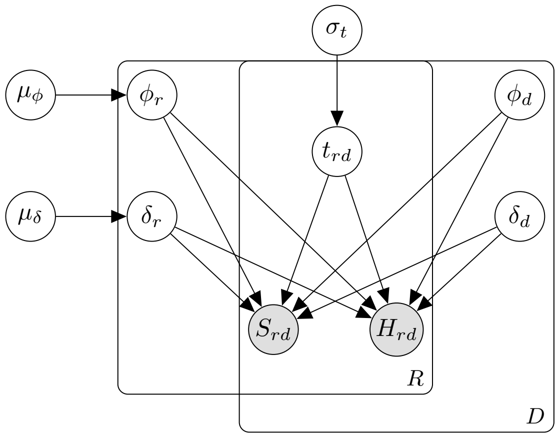

Figure 5. Graphical representation of the generative model for stop decisions. Observed outcomes are shaded, and unshaded nodes are latent variables inferred from data.

4.2.2. RESULTS

We apply the model described above to the subset of approximately 723,000 stops predicated on suspected criminal possession of a weapon, as indicated by officers. In these cases, the stated objective of the stop is discovery of a weapon, and so we consider a stop successful if a weapon was discovered (Goel et al., 2016). To estimate the racial composition of precincts, we use data from the 2010 U.S. Census.

In Figure 6 we compare the inferred stop thresholds for white, black and Hispanic individuals in each precinct. As with frisk decisions, we find stop thresholds for blacks and Hispanics are consistently lower than for whites, suggestive of discrimination. These results are in line with those obtained by the original threshold model (correlation = 0.99). Our findings bolster past studies of New York City's stop-and-frisk practices (Goel et al., 2016; Gelman et al., 2007), which came to similar conclusions, but did not account for potential infra-marginality.

In Table 2 we compare the time to fit our stop model with discriminant distributions rather than beta distributions. As with the frisk model, the speedup from using discriminant distributions is dramatic: with beta distributions, the model requires more than 4 hours to fit; with discriminant distributions it takes under 4 minutes. The primary reason for the speedup is the reduced time per gradient evaluation, although the number of gradient evaluations is also reduced.

A key assumption of our stop model is that the racial composition of the residential population (as estimated by the Census) is similar to the racial composition of pedestrians officers encounter on the street. We test how sensitive our inferred thresholds are to this assumption by refitting the stop model with various estimates of the number of white individuals in each precinct. Specifically, letting $c_{\text{white},d}$ denote the original census estimate, we varied this number from $c_{\text{white},d}/2$ to $2c_{\text{white},d}$. Throughout this range, the inferred thresholds are relatively stable, with thresholds for blacks and Hispanics consistently lower than for whites. This stability is in part due to the fact that altering assumptions about the base population does not change the observed hit rates, which are substantially higher for whites.

put weakly informative priors on ϕ_r, λ_r, and t_{rd}; we put tighter priors on the location parameters ϕ_d and λ_d to restrict geographical heterogeneity and to accelerate convergence.

Table 2. Breakdown of sampling times for the stop model. Each row reports the improvement for the discriminant distribution model relative to the beta distribution model followed by the statistics for both models.

	SPEEDUP	BETA	DISC
SECONDS/n_{EFF}	**77x**	55	0.7
SAMPLES/n_{EFF}	**1x**	18	19
INTEGRATION STEPS/SAMPLE	**2x**	503	220
SECONDS/INTEGRATION STEP	**36x**	.006	.0002
SECONDS TO FIT MODEL	**76x**	15943	211

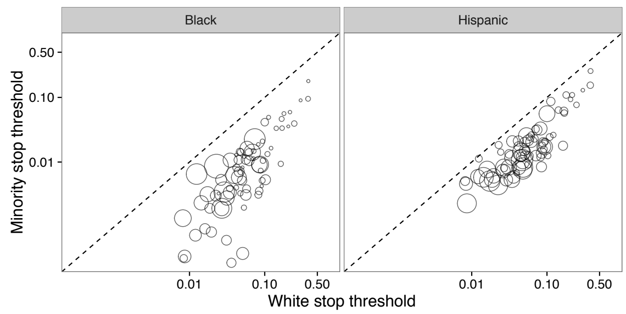

Figure 6. Inferred stop thresholds in the New York City stop-and-frisk data. Axes are logarithmic.

5. Discussion

In this paper we introduced and analyzed discriminant distributions, a two-parameter family on $[0, 1]$ with an intuitive generative interpretation related to linear discriminant analysis. This family of distributions is a restricted subset of 2-component logit-normal mixtures, and as such can closely approximate logit-normal and beta distributions with typical parameter settings. The CDF and conditional mean of discriminant distributions reduce to simple expressions that are no more difficult to evaluate than the equivalent statistics for normal distributions. Because of this simplicity, using discriminant distributions enables us to speed up inference in the threshold test by more than 75-fold.

By cutting inference time from hours to minutes, discriminant distributions make it possible to use the threshold test to investigate discrimination in a wide variety of settings. Practitioners can now iterate over model specifications quickly, facilitating application of the threshold test in domains where infra-marginality is of concern. Some recent examples include medical decisions (Anwar & Fang, 2011) and parole board release decisions (Anwar & Fang, 2015). Our modified test can also scale to much larger datasets, such as the national traffic stop database of Pierson et al. (2017).

Tools for black box Bayesian inference are growing in popularity, with Hamiltonian Monte Carlo and automatic differentiation (Kucukelbir et al., 2016) making it feasible to perform inference for increasingly complicated models. As researchers embrace this complexity, there is opportunity to consider new or uncommon distributions. Historical default distributions—often selected for analytically convenient properties like conjugacy—may not be the best choice when using automatic inference. An early example of this was the Kumaraswamy distribution (Kumaraswamy, 1980; Jones, 2009), developed as an alternative to the beta distribution for its simpler CDF. We believe that discriminant distributions may likewise offer computational speedups beyond the threshold test as automatic inference enjoys increasingly widespread use.

Acknowledgements

We thank Bob Carpenter and Peng Ding for their helpful comments.

References

Anwar, Shamena and Fang, Hanming. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *The American Economic Review*, 96(1):127–151, 2006.

Anwar, Shamena and Fang, Hanming. Testing for the role of prejudice in emergency departments using bounce-back rates. Technical report, National Bureau of Economic Research, 2011.

Anwar, Shamena and Fang, Hanming. Testing for racial prejudice in the parole board release process: Theory and evidence. *The Journal of Legal Studies*, 44(1):1–37, 2015.

Ayres, Ian. Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142, 2002.

Becker, Gary S. *The economics of discrimination*. University of Chicago Press, 1957.

Becker, Gary S. Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, pp. 385–409, 1993.

Betancourt, Michael. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Betancourt, Michael and Girolami, Mark. Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications*, 79:30, 2015.

Bhaduri, Budhendra. Population distribution during the day. In *Encyclopedia of GIS*, pp. 880–885. Springer, 2008.

Boik, Robert J and Robison-Cox, James F. Derivatives of the incomplete beta function. *Journal of Statistical Software*, 3(1):1–20, 1998.

Carpenter, Bob, Gelman, Andrew, Hoffman, Matt, Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Michael A, Guo, Jiqiang, Li, Peter, and Riddell, Allen. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20, 2016.

Chen, Tianqi, Fox, Emily B, and Guestrin, Carlos. Stochastic Gradient Hamiltonian Monte Carlo. In *ICML*, pp. 1683–1691, 2014.

du Plessis, Marthinus C, Niu, Gang, and Sugiyama, Masashi. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems*, pp. 703–711, 2014.

Elkan, Charles and Noto, Keith. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD*, pp. 213–220. ACM, 2008.

Engel, Robin S. A critique of the outcome test in racial profiling research. *Justice Quarterly*, 25(1):1–36, 2008.

Frederic, Patrizio and Lad, Frank. Two moments of the logitnormal distribution. *Communications in Statistics – Simulation and Computation*, 37(7):1263–1269, 2008.

Gelman, Andrew, Meng, Xiao-Li, and Stern, Hal. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pp. 733–760, 1996.

Gelman, Andrew, Fagan, Jeffrey, and Kiss, Alex. An analysis of the New York City police department's stop-and-frisk policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102(479): 813–823, 2007.

Goel, Sharad, Rao, Justin M, and Shroff, Ravi. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016.

Jones, MC. Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology*, 6(1):70–81, 2009.

Kucukelbir, Alp, Tran, Dustin, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*, 2016.

Kumaraswamy, Ponnambalam. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1-2):79–88, 1980.

Mordelet, Fantine and Vert, J-P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

Neal, Radford M et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2:113–162, 2011.

Pierson, Emma, Simoiu, Camelia, Overgoor, Jan, Corbett-Davies, Sam, Ramachandran, Vignesh, Phillips, Cheryl, and Goel, Sharad. A large-scale analysis of racial disparities in police stops across the united states. *Working paper*, 2017.

Simoiu, Camelia, Corbett-Davies, Sam, and Goel, Sharad. The problem of infra-marginality in outcome tests for discrimination. Working paper available at `https://arxiv.org/pdf/1607.05376.pdf`, 2017.

Floyd v. City of New York. 959 F. Supp. 2d 540 - Dist. Court, SD New York. 2013.