

A frequentist local false discovery rate

Daniel Xiang¹, Jake A. Soloff¹, and William Fithian²

¹*Department of Statistics, University of Chicago*

²*Department of Statistics, University of California, Berkeley*

December 12, 2024

Abstract

The local false discovery rate (lfdr) of [Efron et al. \(2001\)](#) enjoys major conceptual and decision-theoretic advantages over the false discovery rate (FDR) as an error criterion in multiple testing, but is only well-defined in Bayesian models where the truth status of each null hypothesis is random. We define a frequentist counterpart to the lfdr based on the relative frequency of nulls at each point in the sample space. The frequentist lfdr is defined without reference to any prior, but preserves several important properties of the Bayesian lfdr: For continuous test statistics, $\text{lfdr}(t)$ gives the probability, conditional on observing *some* statistic equal to t , that the corresponding null hypothesis is true. Evaluating the lfdr at an individual test statistic also yields a calibrated forecast of whether its null hypothesis is true. Finally, thresholding the lfdr at $\frac{1}{1+\lambda}$ gives the best separable rejection rule under the weighted classification loss where Type I errors are λ times as costly as Type II errors. The lfdr can be estimated efficiently using parametric or non-parametric methods, and a closely related error criterion can be provably controlled in finite samples under independence assumptions. Whereas the FDR measures the average quality of all discoveries in a given rejection region, our lfdr measures how the quality of discoveries varies across the rejection region, allowing for a more fine-grained analysis.

1 Introduction

Suppose that we are testing a scientific hypothesis, and observe a z -statistic equal to 3. How confidently can we reject the corresponding null hypothesis in favor of the alternative? This simple and natural question could hardly be better crafted to embarrass frequentist statisticians. Notwithstanding the common misinterpretation of the p -value (in this case roughly 0.0027) as the posterior probability that the null is true in light of the data, calculating this probability in fact requires further information, namely the prior probability that the null is true and the distribution of the test statistic under the alternative. Bayesians are willing to supply these quantities, but face other difficulties: different observers' subjective beliefs may vary widely, and many scientists resist granting that the truth or falsehood of a

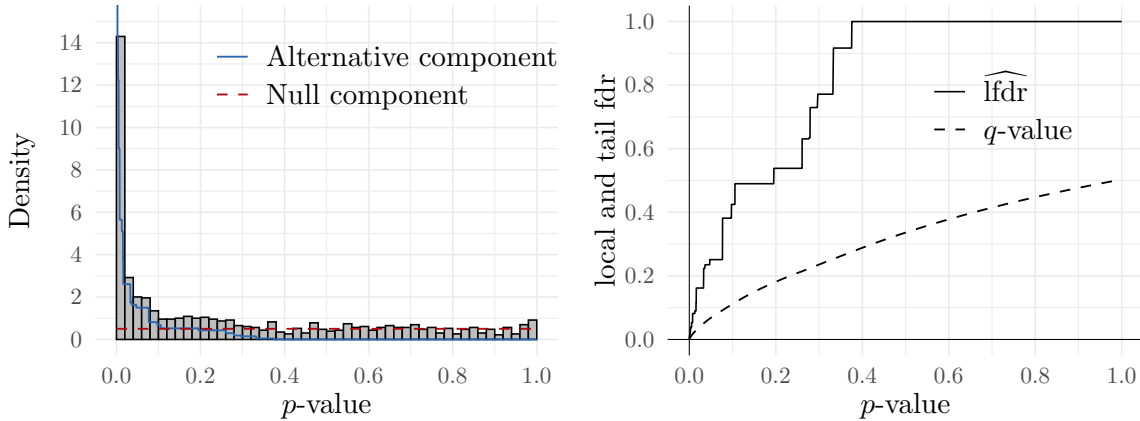


Figure 1: Microbiome preservation example. The left panel shows a histogram of permutation p -values comparing relative abundance in fresh vs. eight-week-old preserved samples in each of $m = 1147$ species, along with a nonparametric estimate of the null and alternative components of the mixture density using the empirical Bayes estimator of [Strimmer \(2008\)](#). The right panel shows the corresponding lfdr estimates, as well as the q -value of [Storey \(2002\)](#). While these lfdr estimates would be difficult to justify as posterior probabilities in a fully Bayesian analysis, they have natural interpretations in our frequentist framework.

concrete scientific hypothesis is a random event whose probability rises and falls according to an observer’s prejudices ([Goodman, 1999](#); [Savage, 1972](#)).

Both frequentists and Bayesians are better positioned to answer the question when the hypothesis is one of many under consideration, provided that the other hypotheses are considered *relevant*, meaning informally that the cases are sufficiently alike to justify a combined analysis. Then, hierarchical or empirical Bayesian methods are appealing because they allow the subjective prior to be replaced with one that is learned from the data. However, calculating posterior probabilities still requires the analyst to model the truth status of individual hypotheses as random variables, and to mathematically formalize the assumption of relevance, typically by assuming that the hypotheses and test statistics are exchangeable across the cases, or else by introducing a parametric model for their dependence. In many scientific contexts, these may be difficult assumptions to accept, even if we are willing in principle to proceed under a Bayesian framework.

Figure 1 shows a histogram of 1147 permutation p -values from a microbiome experiment by [Song et al. \(2016\)](#), discussed in Section 6.2. Each p -value tests whether the relative abundance of a given bacterial species is altered by laboratory storage of biological samples, and is calculated under nonparametric assumptions justified by the controlled experimental conditions. However, the complex taxonomic and ecological relationships between the different species make it highly implausible that the p -values for different species are stochastically independent, or that the scientific hypotheses are exchangeable *a priori*, and the need to model these dependencies presents a forbidding obstacle to carrying out a full Bayesian analysis.

Frequentists reluctant to embrace Bayesian assumptions commonly choose instead to estimate or control the false discovery rate (FDR), defined as the expected fraction of true nulls among hypotheses rejected by a multiple testing method such as the Benjamini–Hochberg (BH) procedure (Benjamini and Hochberg, 1995). The q -values of Storey (2002) give a kind of FDR estimate for individual hypotheses, but they answer a substantively different question than our original one: roughly, the q -value for a hypothesis with test statistic t estimates the proportion of true nulls among all hypotheses with test statistics as extreme *or more extreme* than t . As Figure 1 illustrates, these are two very different questions, and misinterpreting the q -value as a measure of confidence in a given discovery would make us systematically, and often severely, over-optimistic; see Section 3 for further discussion of this point.

Inspired by the FDR, Efron et al. (2001) reinterpreted the BH procedure as an empirical Bayes method, and presented nonparametric methods for estimating the (Bayesian) *local false discovery rate* (lfdr), which they defined as the probability that an individual null hypothesis is true in light of the data; see Section 2.1. Subsequent work has developed practical methods for estimating the lfdr that avoid detailed modeling of dependence wherever possible, but has remained firmly within the paradigm of Bayesian statistics. Thus, the Bayesian lfdr successfully answers our original motivating question, but it does so by reintroducing the prior that the FDR framework so deftly evades.

In this work we introduce a new definition of the lfdr that addresses our motivating question within a fully frequentist model. Suppose that we observe test statistics z_1, \dots, z_m for hypotheses H_1, \dots, H_m , of which m_0 are true nulls. Let $f^{(i)}(t)$ denote the density of z_i , and define the (frequentist) lfdr as the relative frequency of null statistics at each point in the sample space:

$$\text{lfdr}(t) := \sum_{i: H_i \text{ is true}} f^{(i)}(t) \bigg/ \sum_{i=1}^m f^{(i)}(t). \quad (1)$$

If m is large and the dependence is not too strong, $\text{lfdr}(t)$ approximates the proportion of nulls among hypotheses whose statistics fall in a small neighborhood of t .

In the common setting where all of the null statistics share the same density f_0 , we obtain the simpler expression

$$\text{lfdr}(t) = \bar{\pi}_0 f_0(t) / \bar{f}(t), \quad (2)$$

where $\bar{\pi}_0 := m_0/m$ is the true null proportion, and

$$\bar{f}(t) := \frac{1}{m} \sum_{i=1}^m f^{(i)}(t)$$

is the average density. The null density f_0 could, for example, represent the standard Gaussian distribution if the statistics are z -values, or the uniform distribution on $[0, 1]$ if they are p -values. The lfdr estimates in Figure 1 estimate $\bar{\pi}_0$ using Storey’s estimator (Storey, 2002), and $\bar{f}(t)$ using Grenander’s least concave majorant (LCM) estimator (Grenander, 1956). Let \mathcal{Z} generically denote the common sample space where z_1, \dots, z_m are realized. For simplicity of exposition, we will assume throughout that the statistics are continuous, but most of our results extend to discrete sample spaces.

Readers may recognize the expression (2) as nearly identical to the definition of lfdr in the *Bayesian two-groups model* of Efron et al. (2001), but there are key differences. Most importantly, our frequentist lfdr is defined only in terms of the marginal densities of the m statistics, effectively replacing the Bayesian prior with the finite population of cases under study. Correspondingly, our definition of $\text{lfdr}(t)$ is not a Bayesian posterior probability that H_i is true given $z_i = t$, since that probability is either zero or one in our frequentist setting. Instead, the frequentist lfdr is the conditional probability, given *some* statistic equals t , that its null hypothesis is true:

$$\text{lfdr}(t) = \mathbb{P}(H_J \text{ is true} \mid z_J = t, \text{ for some } J). \quad (3)$$

The truth status of H_J in (3) is random because the index J is random.

In Section 2 we prove the relation (3) and establish two other interpretations of our lfdr: First, we show that $\text{lfdr}(z_1), \dots, \text{lfdr}(z_m)$ are calibrated forecasts for the unknown truth of H_1, \dots, H_m . Second, we show that thresholding the lfdr at $1/(1+\lambda)$ implements the optimal decision rule of the form $\delta : \mathcal{Z} \rightarrow \{\text{accept, reject}\}$ for testing H_1, \dots, H_m when Type I errors are λ times as costly as Type II errors, coinciding with the Bayes decision rule in the two-groups model.

Section 3 compares and contrasts the lfdr with the FDR and two related criteria for frequentist error control, the marginal FDR (mFDR) and positive FDR (pFDR). The characterizations of lfdr in (1) and (3) can be understood as limiting forms of the mFDR and pFDR respectively, for a hypothetical rejection rule that “rejects” only statistics in a small neighborhood of t .

Although the frequentist lfdr depends on unknown quantities, it can be estimated efficiently from the data if m is reasonably large and the dependence between test statistics is not too strong, typically using the same methods developed for estimating the Bayesian lfdr. In the formulation (2), f_0 is typically known and $\bar{\pi}_0$ can be conservatively bounded above by 1 or estimated using standard techniques, leaving only the problem of estimating the average density \bar{f} . Section 4 discusses approaches to this problem based on standard parametric or nonparametric methods for density estimation in the i.i.d. setting, and argues that classical empirical Bayes methods commonly applied under the two-groups model can be understood as estimates of the frequentist lfdr.

Section 5 introduces the *boundary FDR*, a frequentist error criterion for multiple testing defined as the probability that the last rejection (i.e., the rejection with the largest p -value) is a false discovery. We show that the support line procedure of Soloff et al. (2024), which is closely related to Strimmer’s monotone lfdr estimator (Strimmer, 2008), controls this criterion in finite samples under an independence assumption. Section 6 illustrates the application of the lfdr in simulation and real data examples, and Section 7 concludes.

2 Interpreting the frequentist lfdr

2.1 Review of the Bayes two-groups model

In prior work, the local false discovery rate has been defined with respect to Bayesian models. The most well-known of these is the so-called *Bayesian two-groups model* of Efron et al. (2001). In that model, each hypothesis has an independent chance π_0 of being true, and the statistic z_i is distributed according to density $f_0(t)$ if H_i is true, and $f_1(t)$ otherwise.

The Bayesian local false discovery rate is defined as the posterior probability that H_i is true in light of the data:

$$\text{lfdr}^*(t) := \mathbb{P}(H_i \text{ is true} \mid z_i = t) = \pi_0 f_0(t) / f(t), \quad (4)$$

where $f(t) = \pi_0 f_0(t) + (1 - \pi_0) f_1(t)$ is the marginal density of z_1, \dots, z_m . To avoid confusion, we use an asterisk to distinguish the Bayesian lfdr from our frequentist lfdr.

When π_0 , f_0 , and f_1 are fixed and known, the posterior probabilities $\text{lfdr}^*(z_1), \dots, \text{lfdr}^*(z_m)$ fully describe the posterior, and they represent the sharpest calibrated forecasts for the truth status of the hypotheses H_1, \dots, H_m (Dawid, 1982; Gupta et al., 2020). They also have a natural interpretation in decision theory. For a decision rule δ that returns an accept/reject decision for each hypothesis, define the *weighted classification loss*, which penalizes the analyst λ for false positives and false negatives at different rates:

$$L_\lambda(H, \delta) := \lambda \cdot (\#\text{false positives}) + (\#\text{false negatives}),$$

where a false positive occurs when δ rejects a true H_i , and a false negative occurs when δ accepts a false H_i . A straightforward calculation shows that the risk-minimizing rule is to reject H_i if and only if $\text{lfdr}^*(z_i)$ is below $1/(1 + \lambda)$ (Sun and Cai, 2007).

This rejection rule is especially simple to interpret when the statistics are p -values with uniform null density $f_0(t) \equiv 1$ on $\mathcal{Z} = [0, 1]$, in which case this rule equates to rejecting H_i whenever z_i is observed in a region with density $f(t) \geq (1 + \lambda)\pi_0$. For example, if a false positive is $\lambda = 4$ times as costly as a false negative, then we should reject when $\text{lfdr}^*(z_i) \leq 0.2$, or equivalently when $f(z_i) \geq 5 \cdot \pi_0$.

The intimate connection between the Bayesian lfdr and the marginal density f has a very convenient consequence in the empirical Bayes setting where f_1 and π_0 are unknown. If $\pi_0 \approx 1$, then estimating the marginal density $f(t)$ from the i.i.d. sample z_1, \dots, z_m is nearly equivalent to estimating $\text{lfdr}^*(t)$, and determining the optimal rejection rule amounts to finding a super-level set of f .

In the two-groups model, these interpretations of the Bayesian lfdr are all easy consequences of standard Bayesian calculations. None of them carry over directly to our frequentist model: if the truth status of H_1, \dots, H_m is fixed, then (i) the probability H_i is true in light of the data is always 0 or 1, (ii) the optimal forecasting rule is to forecast that the true hypotheses are true and the false ones are false, and (iii) the best decision rule for any λ is to reject the false hypotheses and accept the true ones. Nevertheless, all three properties of the Bayesian lfdr have close analogs in our frequentist model, as we explore in the next section.

2.2 Three interpretations of the frequentist lfdr

Section 1 gave three interpretations of the frequentist lfdr that are close analogs of properties enjoyed by the Bayesian lfdr. We now review and elaborate on them:

Interpretation 1: Conditional probability. For a fixed value $t \in \mathcal{Z}$, $\text{lfdr}(t)$ is the conditional probability that a hypothesis with test statistic equal to t is a true null.

Theorem 2.1. *Suppose z_1, \dots, z_m are jointly absolutely continuous. Then*

$$\text{lfdr}(t) = \mathbb{P}(H_J \text{ is true} \mid z_J = t, \text{ for some } J).$$

where J is the (random) index of the statistic with $z_J = t$.

When \mathcal{Z} is discrete (in which case f_0 and f represent probability mass functions) this property does not generalize directly in the way we might initially expect: conditional on the event that *at least one* index J has $z_J = t$, the probability that a randomly selected one is truly null is not in general equal to $\text{lfdr}(t)$, unless we weight the probabilities by the number of statistics equal to t . Instead, we have

$$\text{lfdr}(t) = \frac{\mathbb{E}[\#\{j : z_j = t, H_j \text{ true}\}]}{\mathbb{E}[\#\{j : z_j = t\}]}.$$

This ratio is closely related to the marginal false discovery rate (mFDR). See Section 3 for further discussion of connections between the lfdr, FDR, and mFDR.

Interpretation 2: Calibrated forecast. The lfdr evaluated at the observed statistics z_1, \dots, z_m makes calibrated forecasts for the truth of the null hypotheses H_1, \dots, H_m , where a function $g : \mathbb{R} \rightarrow [0, 1]$ is said to be calibrated if

$$\mathbb{P}(H_J \text{ is true} \mid g(z_J) = \alpha, \text{ for some } J) = \alpha.$$

Theorem 2.2. *Let $\ell_i := \text{lfdr}(z_i)$ for $i = 1, \dots, m$ and suppose z_1, \dots, z_m are jointly absolutely continuous. Then*

$$\mathbb{P}(H_J \text{ is true} \mid \ell_J = \alpha, \text{ for some } J) = \alpha,$$

for any $\alpha \in \text{range}(\text{lfdr})$. Furthermore, lfdr is the finest calibrator in the following sense: if $g : \mathbb{R} \rightarrow [0, 1]$ is calibrated, then for any t ,

$$g(t) = \mathbb{E}(\text{lfdr}(z_I) \mid g(z_I) = g(t)), \tag{5}$$

where $I \sim \text{Uniform}\{1, \dots, m\}$.

Interpretation 3: Optimal rejection rule. Thresholding $\text{lfdr}(z_i)$ at $1/(1 + \lambda)$ gives the optimal separable rejection rule for testing H_1, \dots, H_m under the weighted classification loss with weight λ .

For a decision rule $\delta(z_1, \dots, z_m) \in \{\text{reject}, \text{accept}\}^m$, the weighted classification risk is minimized over separable decision rules by the one that thresholds the frequentist lfdr.

Theorem 2.3. *If δ is a separable decision rule, i.e. $\delta_i(z_1, \dots, z_m) = g(z_i)$ for some univariate function g , then*

$$\mathbb{E}L_\lambda(H, \delta) \geq \mathbb{E}L_\lambda(H, \mathfrak{d}^*),$$

where

$$\mathfrak{d}^*(z_i) = \begin{cases} \text{reject} & \text{if } \text{lfdr}(z_i) \leq \frac{1}{1+\lambda} \\ \text{accept} & \text{otherwise.} \end{cases} \quad (6)$$

2.3 Limitations

The previous section discussed three interpretations of the frequentist lfdr. In practice, lfdr depends on unknown quantities such as $\bar{\pi}_0$ and \bar{f} that must be estimated. Density estimation is a difficult problem in general, and therefore estimating the lfdr can be hard unless we can rely on assumptions like monotonicity or smoothness for the average density \bar{f} . Strong dependence between the test statistics can present an additional complication. We discuss the problem of estimating the lfdr in section 4, and in particular how empirical Bayes estimates target the frequentist lfdr when observations are non-i.i.d.

Another limitation of our lfdr function is that it does not account for additional information that may be known by the analyst. Formally, the conditional probability interpretation applies only to an analyst who is ignorant or indifferent about *which* null hypothesis corresponds to the test statistic realized at, e.g. $z_J = 3$. As a result, the lfdr may not match the posterior belief of an analyst who has different prior opinions about the likelihood of different hypotheses being true. In that case, it may be more appropriate to choose a smaller reference class that represents a subset of the hypotheses under study.

Our last interpretation assumes separability of our decision rule, but this restriction is somewhat artificial. In the absence of covariates, we could instead restrict to permutation equivariant (PE) rules, which implies that the rejection threshold depends only on the set of values $\{z_1, \dots, z_m\}$ and not on the order in which they are observed. In large samples with independent observations, the best PE decision rule is close to the best separable rule, mirroring a well-known phenomenon in the empirical Bayes literature (see, e.g. [Hannan and Robbins \(1955\)](#) and [Greenshtein and Ritov \(2009\)](#)). We elaborate on this point in Section A.1 of the Appendix.

3 lfdr and FDR

The lfdr was originally proposed by [Efron et al. \(2001\)](#) as a modification of the FDR, which measures the expected fraction of Type I errors among all rejections made by a multiple testing procedure. If such a procedure makes $R \geq 0$ rejections (or “discoveries”), of which V correspond to true nulls (or false discoveries), then the *false discovery proportion* (FDP) is defined as the realized fraction of false discoveries, and the FDR is its expectation:

$$\text{FDP} = \frac{V}{1 \vee R}, \quad \text{FDR} = \mathbb{E}[\text{FDP}].$$

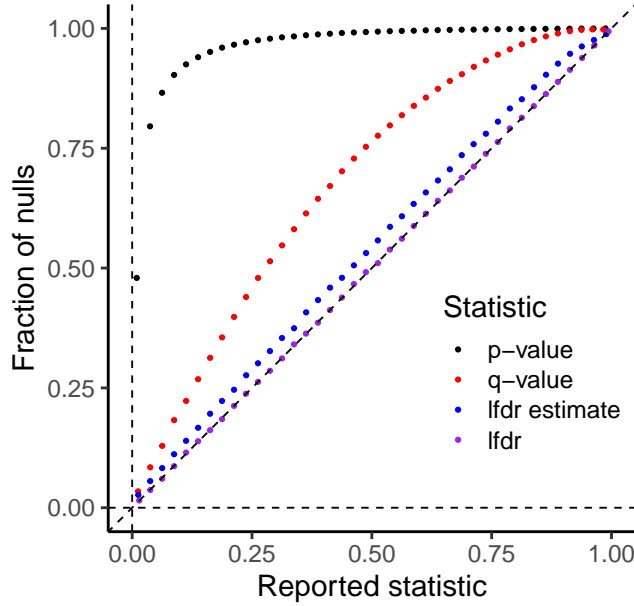


Figure 2: Simulation example. The true means in (8) are $\mu_1 = \dots = \mu_{m_1} = 2$ and $\mu_i = 0$ for $i = m_1 + 1, \dots, m$, where $m = 3000$ and the true null proportion is 95%. The calibration curve for the frequentist lfdr is plotted in purple. We repeat the experiment 10^4 times to assess calibration.

Efron et al. (2001) showed that the BH procedure, the best-known method for controlling the FDR at a pre-specified level q , can be interpreted as an empirical Bayes method that rejects the null for all p -values below a data-adaptive threshold, defined with respect to an estimator of the FDP for all rejection regions of the form $[0, t]$:

$$\hat{t}^{\text{BH}} = \max \left\{ t : \widehat{\text{FDP}}([0, t]) \leq q \right\}, \quad \text{where } \widehat{\text{FDP}}([0, t]) = \frac{mt}{\#\{i : p_i \leq t\}}. \quad (7)$$

When the threshold t is equal to an observation p_i , $\widehat{\text{FDP}}([0, t])$ is the estimated fraction of rejections with evidence at least as significant as p_i that are false discoveries. This quantity is closely related to the q -value (Storey et al., 2004), defined with respect to the BH procedure as follows.

$$\begin{aligned} \hat{q}_{(m)} &= \widehat{\text{FDP}}([0, p_{(m)}]) \\ \hat{q}_{(i)} &= \min \left\{ \widehat{\text{FDP}}([0, p_{(i)}]), \hat{q}_{(i+1)} \right\}, \quad i = m-1, m-2, \dots, 1. \end{aligned}$$

Equivalently, \hat{q}_i is the smallest $q \in [0, 1]$ for which the BH procedure run at level q rejects H_i .

The q -value has a natural Bayesian interpretation as well (Storey, 2003). In the setting of Section 2.1, \hat{q}_i approximates the conditional probability that a null hypothesis is true, given that its p -value fell below the realized value of p_i . As mentioned in Section 1, when we (mistakenly) interpret the q -value as our confidence in a specific null hypothesis conditional on its realized p -value, we tend to be over-confident in our rejections. We demonstrate this point next with a simple numerical experiment.

Consider Gaussian test statistics with known (unit) variance

$$Z_i \sim N(\mu_i, 1) \quad \text{independently for } i = 1, \dots, m = 3000, \quad (8)$$

where $\mu_i \in \{0, 2\}$, and suppose we want to test the null hypotheses $H_i : \mu_i = 0$. From the list of z -statistics, we compute three summary statistics for each null hypothesis: a one-sided p -value, a Storey-BH q -value, and an lfdr estimate based on the ‘`fdrtool`’ package (Strimmer, 2008), except that we use the Storey estimator of π_0 instead of the default estimate. The summary statistics are binned into a grid of $[0, 1]$ with bin size 2.5%, and in each of the forty bins we calculate the proportion of true nulls.

Figure 2 displays the results of the experiment. We see that the q -value systematically under-estimates the chance that the null hypothesis is true, just less extremely than the p -value does. We may reject a null hypothesis H_{ij} at level $q = 25\%$, when our actual confidence in the null hypothesis should be around 50%. By contrast, the true lfdr is exactly calibrated, and the estimated lfdr is nearly so: among t -statistics for which the estimated lfdr is close to 25%, close to a quarter correspond to true null hypotheses.

Two quantities that are closely related to the FDR are the positive FDR and the marginal FDR, defined as

$$\begin{aligned} \text{pFDR}([0, t]) &= \mathbb{E}[\text{FDP}([0, t]) \mid R([0, t]) > 0], \\ \text{mFDR}([0, t]) &= \frac{\mathbb{E}V([0, t])}{\mathbb{E}R([0, t])}, \end{aligned}$$

where $V([0, t]) = \#\{i : p_i \leq t, H_i \text{ is true}\}$ and $R([0, t]) = \#\{i : p_i \leq t\}$. If we shift our perspective from a rejection region $[0, t]$ to a neighborhood $[s, t]$ of the sample space, the corresponding definitions of mFDR and pFDR are

$$\text{mFDR}([s, t]) = \frac{\mathbb{E}V([s, t])}{\mathbb{E}R([s, t])}, \quad \text{pFDR}([s, t]) = \mathbb{E} \left[\frac{V([s, t])}{R([s, t])} \mid R([s, t]) > 0 \right],$$

where $R([s, t])$ and $V([s, t])$ are the number of rejections and Type I errors for the rejection set $[s, t]$. Our frequentist lfdr can be understood as a limiting version of the pFDR and mFDR, provided that the joint density of the test statistics isn’t supported on a zero-Lebesgue-measure subset of \mathbb{R}^m .

Theorem 3.1. *Suppose z_1, \dots, z_m are jointly absolutely continuous. Then, for any t*

$$\lim_{\varepsilon \rightarrow 0} \text{mFDR}([t - \varepsilon, t + \varepsilon]) = \lim_{\varepsilon \rightarrow 0} \text{pFDR}([t - \varepsilon, t + \varepsilon]) = \text{lfdr}(t).$$

The second equality in Theorem 3.1 suggests the following interpretation of the formula for the lfdr: it is roughly equal to the proportion of null hypotheses whose test statistics fell near t . Other conceptions of the lfdr such as (3) and (23) require us to envision independent replications of the entire multiple testing experiment, conditioning on those in which a selected test-statistic is realized near a point. The interpretation in terms of the pFDR demands less from our imagination, making reference to a single collection of hypotheses and their test statistics. The next proposition states that the realized false discovery proportion within a small neighborhood converges to the lfdr in probability as the sample size increases.

Proposition 3.1. *Suppose $z_i \sim f^{(i)}$ are independent, where $f^{(i)} = f_0$ when $H_i = 0$ and $\varepsilon_m \rightarrow 0$ is a sequence with $m_0 \varepsilon_m \rightarrow \infty$, and that $\text{lfd}r : \mathbb{R} \rightarrow [0, 1]$ is differentiable. Let $\alpha \in \text{range}(\text{lfd}r)$. If $|\text{lfd}r'(t)|$ is bounded away from zero on $\{t : \text{lfd}r(t) = \alpha\}$, then*

$$\text{FDP}(\{z_i : |\text{lfd}r(z_i) - \alpha| \leq \varepsilon_m\}) \xrightarrow{\mathbb{P}} \alpha$$

as $m \rightarrow \infty$.

Proposition 3.1 is related to a general calibration theorem in Dawid (1982), who studied forecasting for a binary outcome $Y_m \in \{0, 1\}$ based on observations (Y_1, \dots, Y_{m-1}) . In the current setting, the truth status of the null hypotheses are never revealed, and the forecasts must be made on the basis of the observations z_1, \dots, z_m . Theorem 2.2 implies the $\text{lfd}r$ is exactly calibrated. As Figure 2 demonstrates, estimates of the $\text{lfd}r$ can be approximately calibrated. In the next section, we discuss estimation of the $\text{lfd}r$.

4 Estimating the $\text{lfd}r$

Section 1 expressed the frequentist $\text{lfd}r$ as the intensity ratio $\bar{\pi}_0 f_0(t) / \bar{f}(z)$. Assuming f_0 is known, we may conservatively bound $\bar{\pi}_0 \leq 1$ or estimate it via e.g. Storey's method, reducing the problem of estimating $\text{lfd}r$ to one of estimating the average density \bar{f} .

Closely related is the classical problem of estimating a density f , given i.i.d. observations $z_1, \dots, z_m \sim f$. Many parametric and nonparametric methods have been proposed to estimate f . Consider the maximum likelihood estimator

$$\hat{f}_m := \operatorname{argmax}_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^m \log f(z_i), \quad (9)$$

where \mathcal{F} is a set of candidate density functions.

If f is a monotone (non-increasing) function on $[0, 1]$, which is a common assumption in multiple testing given a sequence of p -values (Genovese and Wasserman (2004), Strimmer (2008)), the method of Grenander (1956) can be used to estimate f using (9) with \mathcal{F} equal to the set of non-increasing probability densities on $[0, 1]$. For Gaussian test-statistics, Kiefer and Wolfowitz (1956) chose \mathcal{F} to be the set of Gaussian mixture densities:

$$\mathcal{F} = \left\{ \int \phi(z - \mu) G(d\mu) : G \text{ is a probability measure} \right\}.$$

Both of these estimators are nonparametric in the sense that the set \mathcal{F} of candidate density functions is infinite-dimensional.

A parametric approach was proposed by Lindsey (1974), where \mathcal{F} is a finite-dimensional exponential family,

$$f(z) = \exp \left\{ \sum_{j=0}^J \beta_j z^j \right\}.$$

The resulting maximum likelihood estimate for f is quite smooth for moderately sized J , e.g. $J = 7$ is the default setting in the ‘`locfdr`’ package of [Efron et al. \(2011\)](#) which implements Lindsey’s method as a sub-routine when estimating the lfdr.

When the observations are not i.i.d., there is no single element of \mathcal{F} for which the objective in (9) matches the log-likelihood of the data. Nevertheless, the M -estimator \hat{f}_m can still be computed from the sequence z_1, \dots, z_m . In the case of Gaussian observations, i.e. $f^{(i)} = N(\theta_i, 1)$, [Zhang \(2009\)](#) argues that it is sensible to estimate the average marginal density \bar{f} using (9), taking \mathcal{F} to be the set of Gaussian location mixture densities. We now restate his intuitive argument in the current setting.

For any candidate function $f \in \mathcal{F}$, the expectation of the objective in (9) is:

$$\begin{aligned} \mathbb{E} \left[\frac{1}{m} \sum_{i=1}^m \log f(z_i) \right] &= \int \frac{1}{m} \sum_{i=1}^m f^{(i)}(z) \log f(z) dz \\ &= \mathbb{E}_{\bar{f}} \log f(Z), \end{aligned}$$

where Z is a draw from the average density \bar{f} . Let f_m^* denote the maximizer for the deterministic analog of (9)

$$\begin{aligned} f_m^* &:= \operatorname{argmax}_{f \in \mathcal{F}} \mathbb{E}_{\bar{f}} \log f(Z) \\ &= \operatorname{argmin}_{f \in \mathcal{F}} D(\bar{f} \parallel f), \end{aligned} \tag{10}$$

where $D(g \parallel h)$ is the KL distance between two probability distributions with densities g and h .

Under sufficient regularity conditions, the maximizer \hat{f}_m will concentrate around f_m^* . In fact, \hat{f}_m will still be a consistent estimate of f_m^* even if the observations are mildly dependent. As long as the objective in (9) converges uniformly to the population-level objective in (10), we will have $D(f_m^* \parallel \hat{f}_m) \xrightarrow{P} 0$ ([Van der Vaart, 2000](#), Theorem 5.7). We record this observation in the following proposition.

Proposition 4.1. *Suppose $M_m^*(f) := \mathbb{E}_{\bar{f}} \log f(Z)$ and $\widehat{M}_m(f) := \frac{1}{m} \sum_{i=1}^m \log f(z_i)$ satisfy*

$$\sup_{f \in \mathcal{F}} |M_m^*(f) - \widehat{M}_m(f)| \xrightarrow{P} 0,$$

as $m \rightarrow \infty$ and suppose that $\bar{f} \in \mathcal{F}$. Then $D(\bar{f} \parallel \hat{f}_m) \xrightarrow{P} 0$.

In situations where $\bar{f} \notin \mathcal{F}$, the M -estimator \hat{f}_m doesn’t target the average density \bar{f} , but instead targets the element of \mathcal{F} that minimizes the KL distance to \bar{f} . To ensure that $f_m^* = \bar{f}$, it is sufficient that each density $f^{(i)}$ belongs to some base class of densities \mathcal{F}_0 , and then we take $\mathcal{F} = \operatorname{conv}(\mathcal{F}_0)$. For example if we knew that each observation was normally distributed with variance 1, then the mixture density \bar{f} is guaranteed to be in the set of Gaussian location mixtures.

Given an estimate of the lfdr, one may use it to perform multiple testing by rejecting all null hypotheses for which this estimate is large. The resulting rejection set aims to control the

lfdr among rejected null hypotheses, but in general comes with no finite-sample guarantees. In the next section, we show that a particular estimate of the lfdr, based on the maximum likelihood estimator of Grenander (1956), leads to a multiple testing procedure of this form that satisfies an exact bound on the false discovery probability of its last rejection. We propose this latter quantity as a new error criterion, called the boundary false discovery rate (bFDR), that is distinct from the usual FDR and more aligned with the concept of lfdr.

5 Controlling the lfdr

To evaluate multiple testing procedures, it is natural to ask whether all the rejections are individually defensible, not just whether the list of all rejections is defensible as a whole. In a Bayesian model, this question can naturally be formulated in terms of the maximum *a posteriori* null probability over all the rejections. Soloff et al. (2024) define the max-lfdr for a multiple testing procedure as the expectation of this maximum, thereby evaluating a procedure $\mathcal{R} = \{i : \text{reject } H_i\}$ according to its *least promising* rejection,

$$\text{max-lfdr}(\mathcal{R}) = \mathbb{E} \left[\max_{i \in \mathcal{R}} \mathbb{P}(H_i \text{ is true} \mid p_i) \right].$$

In a frequentist analysis under the fixed effects model, however, it is less obvious how to formalize what we mean by the “least promising rejection.” In particular, because the null probability for each hypothesis is either one or zero, the maximum is always one whenever we make any false rejections at all.

Instead, we consider the truth status of the null hypothesis associated with the largest p -value within the rejection region. For a procedure \mathcal{R} whose rejection region $[0, \hat{\tau}]$ contains the R smallest p -values, the *boundary false discovery rate* (bFDR) is defined as the probability that at least one rejection was made and the null hypothesis associated with the largest p -value $\leq \hat{\tau}$ is true,

$$\text{bFDR}(\mathcal{R}) := \mathbb{P}(R > 0 \text{ and } H_{(R)} \text{ is true}), \quad (11)$$

where the notation $H_{(k)}$ means the null hypothesis corresponding to the k th smallest p -value. As in (3), the truth status of $H_{(R)}$ is random because the index R is random.

5.1 Comparison with FDR

The usual FDR measures the null probability of a uniformly selected rejection:

$$\text{FDR}(\mathcal{R}) = \mathbb{P}(R > 0 \text{ and } H_{(I)} \text{ is true}), \quad I \sim \text{Uniform}\{1, \dots, R\}.$$

Figure 3 illustrates a numerical example in which the non-null p -values are highly concentrated near zero, leading to a substantial difference between the average-case rejection (FDR), and the ones near the boundary (bFDR).

Under a monotonicity assumption, the boundary rejection has the greatest null probability of any rejection, which implies the boundary FDR is larger than the FDR. While one might

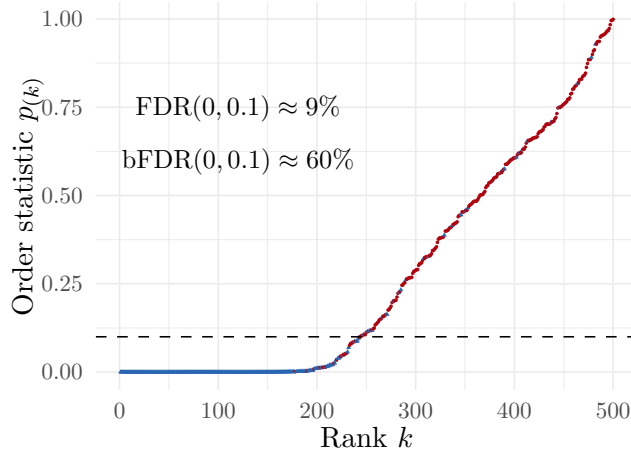


Figure 3: The order statistics of $m = 500$ p -values are plotted against their rank. They are generated with $\bar{\pi}_0 = 0.5$ where null p -values (red) are i.i.d. $\text{Uniform}(0,1)$ and alternative p -values (blue) are i.i.d. $\text{Beta}(0.05,1)$. The bFDR is approximated by the fraction of nulls among the largest 15 p -values below 0.1.

therefore be tempted to conclude that bFDR control is an inherently more conservative goal than FDR control, in practice this may or may not be the case, because one would use a larger threshold when controlling the bFDR than when controlling the FDR. For example, an analyst who equates $\lambda = 4$ type II errors with a single type I error would want to control bFDR at level $1/(1 + \lambda) = 0.2$. The same analyst would *not* be satisfied with a method whose FDR is 0.2, since the cost of the false discoveries would on average exactly cancel out the benefits of the true discoveries.

To illustrate this point, consider the weighted classification risk, which can be redefined (up to additive and multiplicative constants) as

$$L_\lambda(H, \delta) := \lambda V - (R - V),$$

where V is the number of false positives among the R discoveries. Taking $\lambda = 4$, a procedure targeting a false discovery rate $V/R = 1/(1 + \lambda) = 0.2$ achieves the same loss as a trivial procedure that simply sets $V = R = 0$.

Instead, such an analyst would always aim to control FDR at some level smaller than 0.2, for example 0.1 so that they achieve some net benefit from the experiment. As a result, no sensible analyst would ever be interested in bFDR control and FDR control at the same level. Since bFDR control and FDR control typically wouldn't be carried out at the same level, it is unclear which is more conservative in any given case.

5.2 Controlling the boundary FDR

[Soloff et al. \(2024\)](#) proposed the Support Line (SL) method for controlling the max-lfdr under a monotonicity constraint. The procedure run at level α rejects the R_α smallest p -values,

where

$$R_\alpha := \operatorname{argmax}_{k=0,\dots,m} \left\{ \frac{\alpha k}{m} - p_{(k)} \right\}, \quad p_{(0)} := 0. \quad (12)$$

The SL method controls its boundary FDR when the nulls are independent.

Theorem 5.1. *If p_1, \dots, p_m are independent and $p_i \sim \text{Uniform}(0, 1)$ when H_i is true, then*

$$\text{bFDR}(\mathcal{R}_\alpha) = \bar{\pi}_0 \alpha,$$

where $\mathcal{R}_\alpha := \{i : p_i \leq p_{(R_\alpha)}\}$ is defined by (12).

Proof of Theorem 5.1. The event $\{H_{(R_\alpha)} \text{ is true}\}$ can be written as a disjoint union,

$$\{H_{(R_\alpha)} \text{ is true}\} = \bigcup_{i: H_i \text{ is true}} \{p_{(R_\alpha)} = p_i\}$$

which implies

$$\mathbb{P}(H_{(R_\alpha)} \text{ is true}) = \sum_{i: H_i \text{ is true}} \mathbb{P}(p_{(R_\alpha)} = p_i) = m_0 \cdot \frac{\alpha}{m}.$$

The last equality follows from Lemma 2 of [Soloff et al. \(2024\)](#), which states that for any configuration of the other p -values p_1, \dots, p_{m-1} , the probability that a null p -value p_m achieves the optimum in (12) is equal to $\frac{\alpha}{m}$. \square

The proof of Theorem 5.1 shows the boundary FDR control of the SL procedure is controlled when each null density is bounded,

$$H_i \text{ is true} \Rightarrow f^{(i)}(t) \leq 1 \quad \text{for all } t \in [0, \alpha]. \quad (13)$$

This condition is distinct from requiring the nulls be super-uniformly distributed, which is an assumption commonly made in the multiple testing literature and is not sufficient to guarantee boundary FDR control². However, a common example in which super-uniformly distributed p -values arise is one-sided Gaussian location testing,

$$X_i \sim N(\theta_i, 1), \quad i = 1, \dots, m$$

where $H_i \text{ is true} \Rightarrow \theta_i \leq 0$. In this case, the probability density function for $p_i = 1 - \Phi(X_i)$ satisfies (13) under the null, for $\alpha = 1/2$. This observation extends to one-parameter exponential families with continuous densities. The proof of this proposition is recorded in Section B.

¹An alternative proof of the fact “ $p_i \sim \text{Uniform}(0, 1) \Rightarrow \mathbb{P}(p_{(R_\alpha)} = p_i) = \frac{\alpha}{m}$ ” can be found in Section B, the technical key for which is a telescoping sum argument.

²A counterexample to the conjecture that bFDR is controlled by super-uniform p -values is given in Section B.0.2

Proposition 5.1. *Let $(g_\theta)_{\theta \in \mathbb{R}}$ denote an exponential family of continuous distributions on \mathbb{R} with densities*

$$g_\theta(z) = \exp(\theta z - A(\theta))g_{\theta_0}(z), \quad \theta, z \in \mathbb{R},$$

with corresponding cdfs (G_θ) . For one-sided testing of the hypotheses $H_i : \theta_i \leq \theta_0$, let $\alpha^ = 1 - G_{\theta_0}(\mathbb{E}_{\theta_0} Z)$ be the upper quantile of the mean under θ_0 . Then the null density of the one-sided p -value $p = 1 - G_{\theta_0}(Z)$ is bounded by 1 on $[0, \alpha^*]$, for all $\theta \leq \theta_0$.*

Another common setting in which super-uniformly distributed p -values arise is the discrete case, where

$$p_i \sim \text{Uniform}\{1/L, \dots, L/L\} \text{ when } H_i \text{ is true,}$$

for some large fixed grid length L , e.g. the number of permutations used to compute a p -value for a permutation test. In this case, the boundary FDR of the SL procedure is not controlled in finite samples (see Section B.0.1 for a counterexample). However, holding m fixed as $L \rightarrow \infty$, or holding L fixed as $m \rightarrow \infty$, the bFDR of the SL procedure is controlled asymptotically below α . These results are stated precisely in Section A.3, and proved in Section B.

6 Applications

6.1 Example 1: Gaussian graphical model

The Gaussian graphical model is an example of a setting where the frequentist lfdr is useful because the Bayesian approach requires complicated modeling, and the q -value approach is inherently biased.

In this model, the data arrive as n i.i.d. copies of

$$X \sim N(0, \Omega^{-1}),$$

where Ω is a $d \times d$ dimensional precision matrix. This matrix encodes conditional dependence relationships between the coordinates of X as follows: $\Omega_{i,j}$ is zero when the i^{th} and j^{th} coordinate of X are conditionally independent, given the rest of the coordinates. To decide whether or not to reject the null hypothesis:

$$H_{ij} : \Omega_{ij} = 0, \quad i \neq j$$

we may compute a t -statistic on $n - d$ degrees from the linear model obtained by regressing X_j against X_{-j} , taking t_{ij} to be the standardized coefficient for X_i in the fitted model.

For each pair (i, j) , we have $\Omega_{ij} = \Omega_{ji}$ and $t_{ij} = t_{ji}$ so the total number of hypotheses is $\binom{d}{2}$. It would be inappropriate to model the t -statistics as independent, since t_{ij} and t_{jk} being large and positive is informative about the value of t_{ik} . In general, t_{ij} is not a sufficient

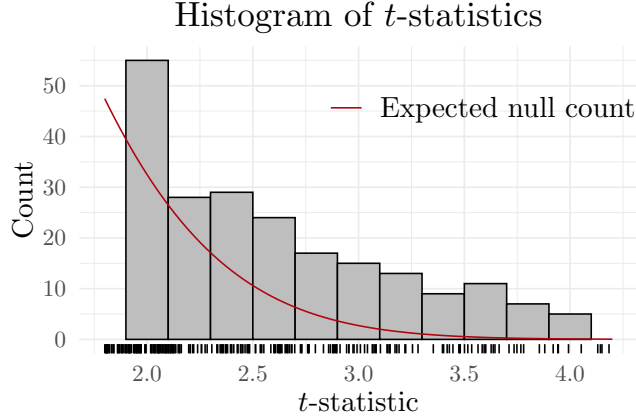


Figure 4: Histogram of t -statistics in the GGM example, with dimensions: $d = 80$, $n = 10d$, $m = 3,160$. The null distribution is t_{n-d} .

statistic for testing H_{ij} , and the posterior probability of the null H_{ij} could be a complicated function of the entire sample covariance matrix.

We can bypass some of these stumbling blocks by calculating a frequentist q -value for each t -statistic, but this can also be misleading. In general, the q -value for t_{ij} substantially underestimates the chance that H_{ij} is true. Figure 4 shows part of the histogram of t -statistics generated from the previously described regression method in a Gaussian graphical model with $d = 80$ and $n = 10d$. Looking at the histogram, it is clear that we can estimate a local null proportion based on the t -statistics. To do so, we first calculate the expected number of null observations at, e.g. $t = 3$. Overlaid in red is the Student- t_{n-d} density weighted by the number of true nulls, in this case $m_0 = 0.95m$. Dividing by the height of the histogram there yields a rough and ready estimate of the lfdr. Compared to the BH q -value, which is around 6% for a t -statistic near 3, the histogram-based estimate of the lfdr is much higher, closer to 20%.

6.2 Example 2: Microbiome data analysis

This section discusses a data set from [Song et al. \(2016\)](#) on storage techniques for biological samples in microbiome analysis. In scientific investigations with microbiome data, it can be necessary to store biological samples for some period of time after collection. A key question about the integrity of the subsequent analysis is whether the relative abundance of different microbial species shifts significantly during the storage period, and whether some storage methods are better than others.

In the data set we analyze, fecal samples from six human participants were stored in 95% ethanol solution, with microbial abundances measured using DNA sequencing techniques, both when the sample was fresh and after eight weeks of storage, with five replications of measurements per participant in each storage condition. In each replication, and for each

of 9719 microbial species³, the relative abundance in each fecal sample is measured as the number of individual bacteria in that species sequenced divided by the total number of bacteria. For each species, we can use a permutation test to assess the null hypothesis that its relative abundance is unchanged after eight weeks of storage.

Due to the careful experimental design, a stratified permutation test is well-suited to test the hypothesis that the relative abundance of a given species is independent of the storage condition (fresh or eight weeks old) given the identity of the human participant. We calculate a p -value p_i for species i based on the Wilcoxon signed rank statistic, where ranks are calculated for the relative abundance of that species within each stratum. Because some species are sparsely observed, we restrict our analysis to the $m = 1147$ species for which the relative abundance is above zero in at least ten total replications. The permutation test is marginally valid for each species, under a generic nonparametric model. Figure 1 shows the p -value histogram as well as a nonparametric estimate for the lfdr and tail FDR, due to [Strimmer \(2008\)](#). To believe in these estimates we need rely only on the assumption that the null p -values are approximately uniform, and that the heights of the histogram bars are informative about the mixture density (or more precisely that the empirical CDF is a good estimator for the true mixture CDF). Both assumptions appear sensible in this case without our needing to appeal to a Bayesian model.

By contrast, it would be highly challenging to specify a convincing Bayesian model for the joint distribution of 9719 species’ relative abundances under the two storage conditions. In particular, given the taxonomic structure of the different species, it is highly unlikely that the true effects of storage on each species are exchangeable, or that the observed relative abundances are independent conditional on the true effects. By shifting to our frequentist perspective, we can sidestep the difficulties of Bayesian modeling.

6.3 Example 3: Aggregate analysis of nudges

The concept of nudging is described by [Thaler and Sunstein \(2009\)](#) as a way of influencing people’s behavior in a predictable way without restricting their options or altering economic incentives. To evaluate the overall effectiveness of psychological nudging on human behavior, [Mertens et al. \(2022a\)](#) collected data from 447 nudge experiments in the behavioral psychology literature. The formulation of this question and the authors’ conclusion was the subject of some debate (see e.g. [Maier et al. \(2022\)](#), [Mertens et al. \(2022b\)](#), [Szasz et al. \(2022\)](#)).

To understand the degree to which false discoveries are present in the aggregated dataset, we estimate the false discovery rate (FDR) using the Storey estimator ([Storey, 2002](#)) for the proportion of true nulls, restricting attention to just the $m = 261$ many p -values falling below the 5% two-sided significance level. This restriction is a way to work around the publication bias present in scientific journals; although ineffective nudges may be under-represented among published studies, the null hypotheses whose p -values fall within the significance region are less prone to censorship ([Hung and Fithian \(2020\)](#), [Jaljuli et al. \(2022\)](#)).

³We use the same method as the original investigators for operationally defining “species,” which are referred to more precisely as *operational taxonomic units* in the microbiome literature.

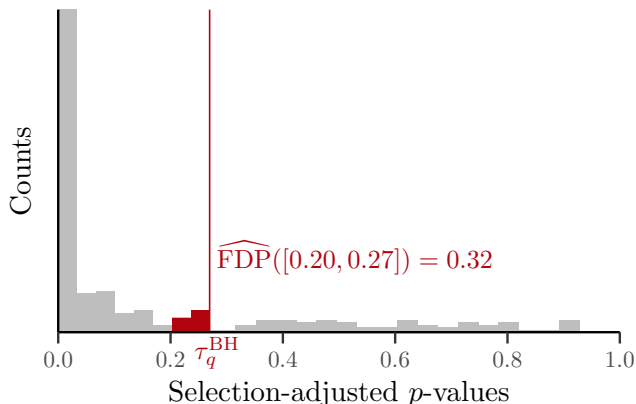


Figure 5: Shown above is the histogram of one-sided p -values falling below 0.025, adjusted for selection by multiplying by 40 (the reciprocal of the 2.5% one-sided significance threshold). The $\text{BH}(q)$ threshold for $q = 10\%$ is around 0.27 (or ≈ 0.007 on the scale of the unadjusted p -values), below which there are 202 rejections. The estimated FDP near the edge of the rejection set (red) is around 32%.

The Storey estimator of the null proportion within the significance region is around 28%, suggesting that roughly a quarter of the $m = 261$ results reported below the 2.5% one-sided significance level are false discoveries. To mitigate the high rate of false claims, we ran the Storey-adjusted BH procedure (Storey et al., 2004) targeting a 10% FDR, yielding a more stringent rejection threshold, as shown in Figure 5, below which there are only 202 p -values.

Upon inspecting the histogram left of the BH threshold, we find that the the estimated rate of false discoveries varies substantially. To estimate the rate of false discoveries within an interval $[s, t]$, we multiplied the total estimated number of nulls by the length of the interval $t - s$, and divided by the total number of observations between s and t . This is a sensible estimate for the mFDR in $[s, t]$ because if the null distribution is $\text{Uniform}(0, 1)$, then $\hat{m}_0(t - s)$ estimates the expected number of false discoveries with p -values between s and t . Visually, this estimate is proportional to the slope of a secant line drawn over the interval $[s, t]$, as illustrated in Figure 6 for the nudge data. As Figure 5 shows, the estimated proportion of false discoveries (FDP) grossly exceeds 10% for a subset of rejections near the rejection threshold.

7 Discussion

“Considering the enormous gains potentially available from empirical Bayes methods, the effects on statistical practice have been somewhat underwhelming.” (Efron, 2019)

One barrier to the wider adoption of empirical Bayes is its philosophical status. Fre-

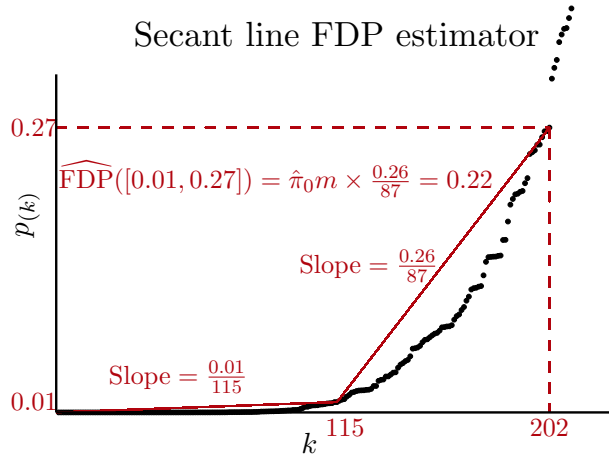


Figure 6: Nudge example. The order statistics of the selection-adjusted p -values are plotted against their rank. The estimated FDP among the smallest 202 p -values is close to 10%. Within the first half, the estimate is 0.6%, whereas in the second half it is 22%.

quantists have legitimate concerns about the Bayesian side of empirical Bayes. This paper introduces a frequentist counterpart to the local false discovery rate: it is firmly rooted in decision theory and enjoys some of the key properties of a Bayesian posterior. Standard empirical Bayes methods estimate this quantity in the frequentist setting.

We close our discussion with some interesting avenues for future work.

- **Frequentist posteriors.** The Bayesian local false discovery rate is simply the posterior of a binary latent variable, H_i , and our frequentist definition corresponds to the oracle Bayes posterior of [Efron \(2019\)](#). It may be of interest to estimate the full oracle Bayes posterior beyond binary settings. In the Gaussian sequence model, compound decision theory has mostly focused on estimating the mean of the posterior ([Zhang, 2009](#); [Jiang and Zhang, 2009](#)).
- **Estimation in the frequentist model.** While we give one asymptotic result on estimating the lfdr (Proposition 4.1), finite-sample estimation error is a serious concern. When the statistics are independent but not identically distributed, the empirical distribution is in a strong sense less dispersed than its i.i.d. counterpart (see, e.g., [Shorack and Wellner, 2009](#), Chapter 25). Does this observation allow us to translate empirical Bayes guarantees into compound decision theory guarantees ([Hannan and Robbins, 1955](#); [Han and Niles-Weed, 2024](#))? How robust are empirical Bayes estimates of marginal lfdr to violations of independence?

Acknowledgements

This project was originally inspired by a question posed by Brad Efron after a seminar in 2018: what, if anything, does the lfdr mean if we do not believe in our Bayesian assumptions? A frequentist lfdr also appeared in an earlier version of [Weinstein et al. \(2019\)](#), which further inspired us to continue developing the frequentist interpretation of lfdr. We thank Bradley Efron, Rina Foygel Barber, Chao Gao, Ruth Heller, Peter McCullagh, Asaf Weinstein, Dani Yekutieli, and Stefan Wager for helpful comments, discussions and support. J. A. Soloff gratefully acknowledges the support of the National Science Foundation via grant DMS-2023109, the Office of Naval Research via grant N00014-20-1-2337, and the Margot and Tom Pritzker Foundation.

Reproducibility. Code to reproduce all figures is available at our [Github repository](#)⁴.

A Appendix

A.1 Compound lfdr

In this section, we refer to formula (2) as the marginal lfdr since it scores the i th null hypothesis as a function of only its p -value p_i . In practice, we would need to estimate the quantities $\bar{\pi}_0, \bar{f}_0, \bar{f}$ appearing in (2), so our decision to reject or accept the i th null hypothesis eventually depends on all of p_1, \dots, p_m . In the absence of further contextual information, it is natural to require the decision rule to be symmetric with respect to the order in which the p -values are observed. This symmetry elicits another oracle function, called the compound lfdr, which plays a role parallel to that of the lfdr in characterizing the best permutation equivariant decision rule.

We say that a decision rule $\delta(\mathbf{p}) := (\delta_1(\mathbf{p}), \dots, \delta_m(\mathbf{p}))$ is permutation equivariant (PE) if

$$\delta(\mathbf{p})_\pi = \delta(\mathbf{p}_\pi) \quad \text{for any } \pi \in \mathcal{S}_m, \quad (14)$$

where \mathcal{S}_m is the set of permutations on $[m]$, and $v_\pi := (v_{\pi(1)}, \dots, v_{\pi(m)})$ denotes the vector $v \in \mathbb{R}^m$ permuted by π . Any multiple testing procedure that uses a rejection threshold which is a function of the order statistics is PE. For example, the Benjamini-Hochberg procedure ([Benjamini and Hochberg, 1995](#)) applied to a list of p -values defines a PE decision rule.

Random shuffling induces an exchangeable Bayesian model:

$$\begin{aligned} \pi &\sim \text{Uniform}(\mathcal{S}_m) \\ \tilde{H} &:= H_\pi, \quad \tilde{\mathbf{p}} := \mathbf{p}_\pi. \end{aligned} \quad (15)$$

The weighted classification risk of any PE decision rule δ in this model coincides with its frequentist compound risk, yielding another instance of the fundamental theorem for compound

⁴full link: <https://github.com/dan-xiang/dan-xiang.github.io/tree/master/frequentist-lfdr-paper>

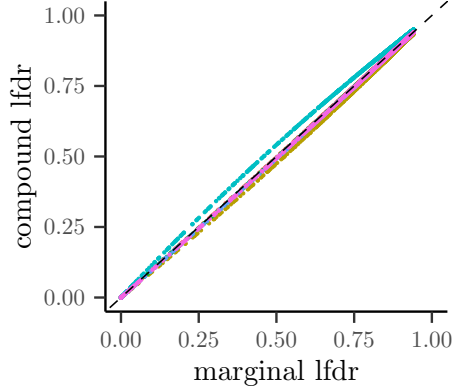


Figure 7: For each of 6 realizations of the vector (p_1, \dots, p_m) , with $m = 1000$, $\bar{\pi}_0 = 0.8$, $f_0 = 1_{[0,1]}$ and $f_1 = \text{Beta}(1/4, 1)$, $\text{clfdr}(\mathbf{p})$ is approximated numerically and the points $(\text{lfdr}(p_i), \text{clfdr}_i(\mathbf{p}))$ are plotted with the diagonal $y = x$ shown as a dashed line. Each color represents a different realization of the one-thousand p -values.

decisions (Zhang (2003), Robbins (1951), Weinstein (2021))

$$\mathbb{E}L_\lambda(H, \delta(\mathbf{p})) = \tilde{\mathbb{E}}L_\lambda(\tilde{H}, \delta(\tilde{\mathbf{p}})),$$

where $\tilde{\mathbb{E}}$ marginalizes over \tilde{H} and $\tilde{\mathbf{p}}$ generated by (15). The right hand side is minimized by the Bayes rule within the exchangeable oracle model (15), characterized by the compound lfdr (clfdr),

$$\begin{aligned} \text{clfdr}_i(t_1, \dots, t_m) &:= \mathbb{P}(\tilde{H}_i \text{ is true} \mid \tilde{\mathbf{p}} = \mathbf{t}) \\ &= \frac{\sum_{\pi \in \mathcal{S}_m: H_{\pi(i)} \text{ is true}} \prod_{j=1}^m f^{(\pi(j))}(t_j)}{\sum_{\pi \in \mathcal{S}_m} \prod_{j=1}^m f^{(\pi(j))}(t_j)}, \end{aligned} \quad (16)$$

for $i = 1, \dots, m$ and $\mathbf{t} := (t_1, \dots, t_m) \in [0, 1]^m$. It follows that the best PE decision rule is

$$\delta_i^*(\mathbf{p}) := \begin{cases} 1 & \text{if } \text{clfdr}_i(\mathbf{p}) \leq \frac{1}{1+\lambda} \\ 0 & \text{else.} \end{cases} \quad (17)$$

This claim follows from a more general relationship between the best PE decision rule and the Bayes rule with respect to a Haar measure prior (see Eaton and George (2021) for a paraphrasing of this result). We also include an elementary proof in appendix B for completeness.

The marginal lfdr is recovered in the exchangeable model (15) by conditioning on one p -value,

$$\text{lfdr}(t) = \mathbb{P}(\tilde{H}_i \text{ is true} \mid \tilde{p}_i = t), \quad t \in [0, 1].$$

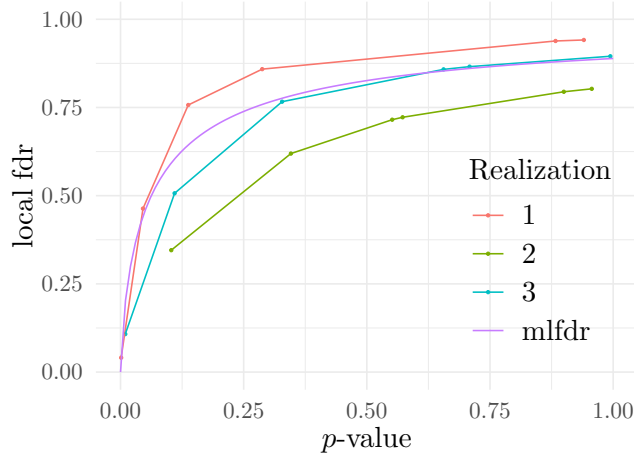


Figure 8: Three realizations of the order statistics of $m = 6$ many p -values are plotted against their scores $\text{clfdr}_i(\mathbf{p})$ for $i = 1, \dots, 6$. In this simulation, $m_0 = 4$, $f_0 = \text{Uniform}(0, 1)$, and $f_1 = \text{Beta}(1/4, 1)$. The clfdr scores are computed using the realized values $p_{(1)} \leq \dots \leq p_{(6)}$ and formula (16). The dependence between clfdr scores in any given realization requires that they always sum to m_0 .

Given the true p -value densities $f^{(1)}, \dots, f^{(m)}$, the clfdr can typically only be computed in small problems (e.g. $m \leq 20$), but can be approximated numerically in larger problems (e.g. $m \approx 1000$) using a method developed by McCullagh (2014) for approximating a matrix permanent. Whereas the lfd r is a fixed function on $[0, 1]$, clfdr depends on the particular realization of p -values, as illustrated in Figure 8. The clfdr and lfd r scores are plotted for six realizations of p -values in Figure 7, where they can be seen to roughly coincide for large m .

In the next section, we discuss the marginal and compound lfd r functions from a Bayesian perspective. Bayesians with an exchangeable prior implicitly report their estimate of the clfdr via their posterior null probability given all the observations. In light of the previous discussion, this implies that the marginal lfd r is close to the “right” answer in any Bayesian model where the prior is exchangeable and the observations are independent given the truth status of each null hypothesis.

A.2 Bayesian interpretation of clfdr and mlfdr

In a Bayesian model, $\text{clfdr}_i(\mathbf{p})$ is the conditional probability that the i th null hypothesis is true, given the data and the empirical distribution of the underlying parameters. For context, suppose there is an exchangeable sequence of latent variables $\theta_1, \dots, \theta_m$ taking values in some parameter space Θ , and conditional on $\theta = (\theta_1, \dots, \theta_m)$, the data is drawn according to

$$p_i \mid \theta \sim f_{\theta_i}, \quad \text{independently for } i = 1, \dots, m. \quad (18)$$

A standard example is the normal location model, where $f_{\theta_i}(\bar{\Phi}^{-1}(p_i)) = \phi(\bar{\Phi}^{-1}(p_i) - \theta_i)$ and ϕ is the standard normal density. The more general setting is recovered by taking the parameter

to be $\theta_i = (H_i, f^{(i)})$ and the parameter space to be $\Theta = \{0, 1\} \times \{\text{all densities on } [0, 1]\}$.

For a given realization of θ , the marginal and compound lfdr in a Bayesian model with an exchangeable prior on θ are:

$$\text{lfdr}(t; G_m) = \mathbb{P}(\theta_i = 0 \mid p_i = t, G_m), \quad (19)$$

$$\text{clfdr}_i(\mathbf{t}; G_m) = \mathbb{P}(\theta_i = 0 \mid \mathbf{p} = \mathbf{t}, G_m) \quad (20)$$

where $t \in [0, 1]$, $\mathbf{t} \in [0, 1]^m$,

$$G_m(t) := m^{-1} \sum_{i=1}^m 1\{\theta_i \leq t\}$$

is the empirical cumulative distribution function of the true effects, and $\{\theta_i = 0\}$ is the null event⁵. This definition appears ambiguous, because up until this point, the marginal and compound lfdr have only been defined in a strictly frequentist model. To clarify, conditioned on a specific realization of θ , the joint distribution (18) defines a frequentist model, and within this model the frequentist lfdr and clfdr functions are equivalent to posterior probabilities that condition also on G_m within the ambient Bayesian model.

Proposition A.1. *Suppose $\theta_1, \dots, \theta_m$ is an exchangeable sequence of latent variables, and that given $\theta = (\theta_1, \dots, \theta_m)$, the p -values are drawn according to (18). Let $\text{lfdr}(t; G_m)$ and $\text{clfdr}_i(\mathbf{t}; G_m)$ be defined as in (19) and (20) where $\mathbf{t} := (t_1, \dots, t_m) \in [0, 1]^m$. Then*

$$\begin{aligned} \text{lfdr}(t; G_m) &= \frac{\bar{\pi}_0 f_0(t)}{\frac{1}{m} \sum_{i=1}^m f_{\theta_i}(t)} \\ \text{clfdr}_i(\mathbf{t}; G_m) &= \frac{\sum_{\pi \in \mathcal{S}_m: \theta_{\pi(i)}=0} \prod_{j=1}^m f_{\theta_{\pi(j)}}(t_j)}{\sum_{\pi \in \mathcal{S}_m} \prod_{j=1}^m f_{\theta_{\pi(j)}}(t_j)}, \end{aligned}$$

for $i = 1, \dots, m$, where $\bar{\pi}_0 := \frac{\#\{i: \theta_i=0\}}{m}$.

For the compound lfdr, there is a large class of Bayesians (essentially, ones with exchangeable priors over $(\theta_i)_{i=1}^m$) for whom their posterior credence in each null hypothesis coincides with their Bayes estimate of compound lfdr. In this sense, we might say Bayesians with exchangeable priors are all in agreement that the compound lfdr is the right quantity to estimate. The same can nearly be said about the marginal lfdr, for the smaller subclass of Bayesians who look marginally at the data for each hypothesis. For these Bayesians, the posterior probability given a single p_i coincides with their conditional expectation of the lfdr. These claims are formalized in the next proposition, which is a straightforward consequence of the tower property of conditional expectations.

Proposition A.2. *Suppose the sequence $\{(\theta_i, p_i)\}_{i=1}^m$ is exchangeable and (18) holds for each $i = 1, \dots, m$. Then*

$$\mathbb{P}(\theta_i = 0 \mid \mathbf{p}) = \mathbb{E}[\text{clfdr}_i(\mathbf{p}; \theta) \mid \mathbf{p}]. \quad (21)$$

⁵previously denoted “ H_i is true”

Marginally, we have for each $i = 1, \dots, m$

$$\mathbb{P}(\theta_i = 0 \mid p_i) = \mathbb{E}[\text{lfdr}(p_i; \theta) \mid p_i]. \quad (22)$$

If we can obtain a good estimator of the compound lfdr given structural assumptions like monotonicity, then any Bayesian with an exchangeable prior on the hypotheses should be fairly satisfied with using it to make predictions, since the predictions they would make are just their estimate of the same quantity. In particular, in many large problems, most of these Bayesian observers would converge on similar estimates for compound lfdr. In such cases, a good frequentist estimator of compound lfdr should also give about the same answer.

The marginal lfdr is computationally simpler to evaluate than the compound lfdr, and under sufficiently regular conditions, their ratio tends to 1 as $m \rightarrow \infty$.

Lemma A.1. *Suppose $p_i \sim f^{(i)}$ are drawn independently for $i = 1, \dots, m$ where each $f^{(i)}$ is a continuous density. $f^{(i)} = f_0$ when $H_i = 0$ and $f^{(i)} = f_1$ when $H_i = 1$. If $\frac{m_0}{m} \rightarrow \pi_0 \in (0, 1)$ as $m \rightarrow \infty$, and $\text{Var}\left(\frac{f_1}{f_0}(p_1)\right) \vee \text{Var}\left(\frac{f_0}{f_1}(p_2)\right) < \infty$ when $p_1 \sim f_0$ and $p_2 \sim f_1$, then we have for each $i = 1, 2, \dots$*

$$\mathbb{P}\left(\left|\frac{\text{clfdr}_i(\mathbf{p})}{\text{lfdr}(p_i)} - 1\right| > m^{-1/2}(\log m)^{3/2}\right) \leq \frac{C}{(\log m)^3}$$

for some constant $C > 0$ when m is sufficiently large.

A.3 Discrete-uniform null distribution

The results that follow in this section are asymptotic, and regard the boundary FDR of the SL procedure run directly on the discrete p -values. If either $L \rightarrow \infty$ with m fixed, or $m \rightarrow \infty$ with L fixed, then the boundary FDR is controlled asymptotically (Theorems A.1 and A.2). Things are less clear when m/L converges to a constant. In this case, numerical evidence suggests the bFDR guarantee may be violated even as m, L get large, with $L \approx 1.8m$ (Figure 9). Lemma A.2 sheds some light on the interplay between m and L and is important for proving our next result.

Theorem A.1. *Let p_1, \dots, p_m be independent random variables with the same support, where $p_i \sim \text{Uniform}\left\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{L-1}{L}, 1\right\}$ when H_i is true. For any fixed $\alpha \in [0, 1]$ and m , we have*

$$\text{bFDR}(\mathcal{R}_\alpha) \leq \bar{\pi}_0 \alpha + O\left(\frac{m^2}{L}\right),$$

as $L \rightarrow \infty$.

The proof of Theorem A.1 is similar to that of Theorem 5.1, both of which start by splitting the boundary FDR into a symmetric contribution from each null,

$$\text{bFDR}(\mathcal{R}_\alpha) = m_0 \mathbb{P}(\text{rank}(p_m) = R),$$

breaking ties uniformly at random, and assuming without loss of generality that $H_m = 0$. As $L \rightarrow \infty$, the probability that p_m is the realized boundary rejection is asymptotically no larger than α/m .

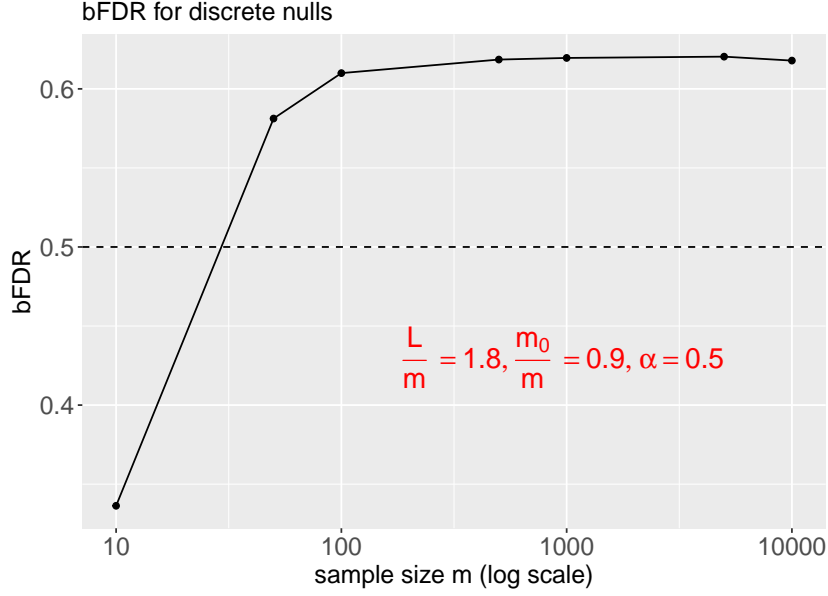


Figure 9: The simulation setting is $\bar{\pi}_0 = 0.9$, $\alpha = 0.5$, and $L = 1.8m$, and the non-nulls are fixed along the grid $\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{m_1}{L}\}$. The sample size ranges from $m \in \{10, 50, 100, 500, 1000, 5000, 10^4\}$. The bFDR of the SL procedure is estimated using $N = 10^5$ Monte Carlo samples.

Lemma A.2. *Let $p_1, \dots, p_{m-1} \in [0, 1]$ be deterministic (non-random) variables, and suppose that $p_m \sim \text{Uniform}\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{L-1}{L}, 1\}$. Then as $L \rightarrow \infty$,*

$$\mathbb{P}(\text{rank}(p_m) = R) \leq \frac{\alpha}{m} + O\left(\frac{m}{L}\right),$$

as $L \rightarrow \infty$, where $\text{rank}(p_m) := \#\{i : p_i \leq p_m\}$ is the rank of p_m among the full list p_1, \dots, p_m , breaking ties at random.

The assumption that $L \rightarrow \infty$ in the previous lemma is needed to rule out small cases where it is possible for the inequality to be violated⁶ by a factor of 2. Our next result characterizes the boundary FDR of the SL procedure as $m \rightarrow \infty$, keeping L fixed.

Theorem A.2. *Let p_1, \dots, p_m be independent random variables with the same support, where $p_i \sim \text{Uniform}\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{L-1}{L}, 1\}$ when H_i is true, for some fixed L . Suppose as $m \rightarrow \infty$ that the average probability mass function \bar{f} converges to a limiting pmf f^* , supported on $\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{L-1}{L}, 1\}$, and that $\frac{m_0}{m} \rightarrow \pi_0^* \in (0, 1)$. Further assume that there is a unique maximizer ℓ^* of the population-level objective,*

$$\ell^* := \operatorname{argmax}_{\ell=0, \dots, L} \left\{ \alpha \sum_{k=0}^{\ell} f^*(k/L) - \ell/L \right\}.$$

⁶A counterexample is obtained by setting p_1, \dots, p_{m-1} to specific values for which $\mathbb{P}(p_{(R)} = p_m) > \frac{2\alpha}{m}$ when $L = 9, m = 6, \alpha = 1/2$. More details about this counter-example can be found in Section B.0.1.

Then we have

$$\lim_{m \rightarrow \infty} \text{bFDR}(\mathcal{R}_\alpha) = \frac{\pi_0^*}{L f^*(\ell^*/L)} \cdot 1_{\{\ell^* > 0\}}.$$

Remark A.1. It follows as a corollary of Theorem A.2 that $\text{bFDR}(\mathcal{R}_\alpha) \leq \pi_0^* \alpha$ asymptotically (as $m \rightarrow \infty$, keeping L fixed) because ℓ^* must occur at an $\ell \leq L$ for which the discrete difference sequence is non-negative,

$$\alpha f^*(\ell^*/L) - 1/L \geq 0,$$

which implies that $\frac{\pi_0^*/L}{f^*(\ell^*/L)} \leq \pi_0^* \alpha$.

B Proofs of technical results

Proof of Theorem 2.1. Given that some $z_J = t$, the index J is a random variable satisfying

$$\mathbb{P}(J = j \mid z_J = t) \propto f^{(j)}(t).$$

By the continuous density assumption, J is almost surely unique. Therefore,

$$\mathbb{P}(J \in \mathcal{H}_0 \mid z_J = t) = \frac{\sum_{j \in \mathcal{H}_0} f^{(j)}(t)}{\sum_{j=1}^m f^{(j)}(t)} = \frac{\bar{\pi}_0 f_0(t)}{\bar{f}(t)}. \quad \square$$

Proof of Theorem 2.2. The continuous density assumption implies

$$\begin{aligned} \mathbb{P}\left(\bigcup_{j \in \mathcal{H}_0} \{|\ell_j - \alpha| \leq \varepsilon\}\right) &\sim \sum_{j \in \mathcal{H}_0} \int_{\{t: |\text{lfd}_j(t) - \alpha| \leq \varepsilon\}} f_0(t) dt \\ \mathbb{P}\left(\bigcup_{j=1}^m \{|\ell_j - \alpha| \leq \varepsilon\}\right) &\sim m \int_{\{t: |\text{lfd}(t) - \alpha| \leq \varepsilon\}} \bar{f}(t) dt, \end{aligned}$$

as $\varepsilon \rightarrow 0$. Now since $\bar{\pi}_0 f_0(t) \sim \alpha \bar{f}(t)$ for any t such that $|\text{lfd}(t) - \alpha| \leq \varepsilon$, the ratio tends to α .

Since g is calibrated,

$$\begin{aligned} g(t) &= \mathbb{E}[1\{H_I \text{ is true}\} \mid g(z_I) = g(t)] \\ &= \mathbb{E}[\text{lfd}_I(z_I) \mid g(z_I) = g(t)], \end{aligned}$$

by the tower property, where $I \sim \text{Uniform}\{1, \dots, m\}$. \square

Proof of Theorem 2.3. The expected weighted classification loss can be re-expressed

$$\mathbb{E} L_\lambda(H, \delta) = \mathbb{E}_{z_I \sim \bar{f}} \ell_\lambda(H_I, g(z_I)),$$

where $I \sim \text{Uniform}\{1, \dots, m\}$ and $\ell_\lambda(h, y)$ is the per-instance loss:

$$\begin{aligned} \ell_\lambda(h, y) &= \lambda 1\{h \text{ is true}, y = \text{reject}\} \\ &\quad + 1\{h \text{ is false}, y = \text{accept}\}. \end{aligned}$$

Since $z_I \sim \bar{f} = \bar{\pi}_0 f_0 + (1 - \bar{\pi}_0) \bar{f}_1$ follows a Bayesian two-groups model, the expected loss is minimized by the Bayes rule (Sun and Cai, 2007), which is characterized by the local fdr in this two-groups model,

$$\text{lfdr}(t) = \mathbb{P}(H_I \text{ is true} \mid z_I = t). \quad (23)$$

□

Proof of Theorem 3.1. Let $V \equiv V(z - \varepsilon, z + \varepsilon)$ and $R \equiv R(z - \varepsilon, z + \varepsilon)$. First, $\text{mFDR}([t - \varepsilon, t + \varepsilon])$ is equal to

$$\begin{aligned} &= \frac{\mathbb{E}[\sum_{i=1}^m (1 - H_i) \mathbf{1}\{z_i \in [t - \varepsilon, t + \varepsilon]\}]}{\mathbb{E}[\sum_{i=1}^m \mathbf{1}\{z_i \in [t - \varepsilon, t + \varepsilon]\}]} \\ &= \frac{\sum_{i=1}^m (1 - H_i) \mathbb{P}(z_i \in [t - \varepsilon, t + \varepsilon])}{\sum_{i=1}^m \mathbb{P}(z_i \in [t - \varepsilon, t + \varepsilon])} \\ &\sim \frac{\sum_{i: H_i=0} f^{(i)}(t) \varepsilon}{\sum_{i=1}^m f^{(i)}(t) \varepsilon} = \text{lfdr}(t). \end{aligned}$$

The third line holds because each density $f^{(i)}$ is continuous, so the probability that $z_i \in [t - \varepsilon, t + \varepsilon]$ is proportional to the density evaluated at some point in this interval, multiplied by the length of the interval.

Next, the pFDR is

$$\mathbb{E}(V/R \mid R > 0) = \frac{\mathbb{E}(V/R \cdot \mathbf{1}\{R > 0\})}{\mathbb{P}(R > 0)}.$$

Since $\mathbb{P}(R = 1 \mid R > 0) \rightarrow 1$ as $\varepsilon \rightarrow 0$, we have

$$\mathbb{E}(V/R \cdot \mathbf{1}\{R > 0\}) \sim \mathbb{P}(\cup_{j \in \mathcal{H}_0} \{|Z_j - z| < \varepsilon\}).$$

Since $\mathbb{P}(R > 0) = \mathbb{P}(\cup_{j \in [m]} \{|Z_j - z| < \varepsilon\})$, the ratio is

$$\begin{aligned} &= \lim_{\varepsilon \rightarrow 0} \frac{\mathbb{P}(\cup_{j \in \mathcal{H}_0} \{|Z_j - z| < \varepsilon\})}{\mathbb{P}(\cup_{j \in [m]} \{|Z_j - z| < \varepsilon\})} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{\sum_{j \in \mathcal{H}_0} \mathbb{P}\{|Z_j - z| < \varepsilon\} + O(\varepsilon^2)}{\sum_{j \in [m]} \mathbb{P}\{|Z_j - z| < \varepsilon\} + O(\varepsilon^2)} = \text{lfdr}(z). \end{aligned}$$

□

Proof of Proposition 3.1. Let $S_m := \{t \in [0, 1] : |\text{lfdr}(t) - \alpha| \leq \varepsilon_m\}$, and define the count variables

$$N_0 := \#\{i \in \mathcal{H}_0 : p_i \in S_m\}, \quad N_1 := \#\{i \in \mathcal{H}_1 : p_i \in S_m\}.$$

To show that $\frac{N_0}{1 \vee (N_0 + N_1)} - \alpha \rightarrow 0$ in probability, it is enough to show

$$\frac{N_0}{\mathbb{E}N_0} - 1 \xrightarrow{\mathbb{P}} 0, \quad \text{and} \quad \frac{N_0 + N_1}{\mathbb{E}(N_0 + N_1)} - 1 \xrightarrow{\mathbb{P}} 0, \quad (24)$$

as $m \rightarrow \infty$, since the ratio of expectations $\frac{\mathbb{E}(N_0)}{\mathbb{E}(N_0 + N_1)} \rightarrow \text{lfdr}(\text{lfdr}^{-1}(\alpha)) = \alpha$. By Chebyshev's inequality, we have for any fixed $\delta > 0$,

$$\mathbb{P}\left(\left|\frac{N_0}{\mathbb{E}N_0} - 1\right| > \delta\right) \leq \frac{\text{Var}(N_0)}{(\mathbb{E}N_0)^2\delta^2} \asymp \frac{m_0\varepsilon_m}{(m_0\varepsilon_m)^2}, \quad (25)$$

because $H_i = 0$ implies $\mathbb{P}(\text{lfdr}(z_i) \in \alpha \pm \varepsilon_m) \asymp \varepsilon_m$ as $m \rightarrow \infty$, since for each α in the range of lfdr , the absolute derivative $|\text{lfdr}'(t)|$ is bounded away from zero on $\{t : \text{lfdr}(t) = \alpha\}$. The right hand side of (25) tends to zero since $m_0\varepsilon_m \rightarrow \infty$. Similarly,

$$\begin{aligned} \mathbb{P}\left(\left|\frac{N_0 + N_1}{\mathbb{E}(N_0 + N_1)} - 1\right| > \delta\right) &\leq \frac{\text{Var}(N_0 + N_1)}{(\mathbb{E}(N_0 + N_1))^2\delta^2} \\ &\asymp \frac{m\varepsilon_m}{(m\varepsilon_m)^2} \leq \frac{1}{m_0\varepsilon_m} \rightarrow 0, \end{aligned}$$

from which (24) follows. \square

Proof of Proposition 4.1. By Theorem 5.7 in Van der Vaart (2000), it suffices to check that f_m^* is well-separated, i.e. for every $\varepsilon > 0$,

$$\sup_{f \in \mathcal{F}: D(f_m^* \| f) \geq \varepsilon} M_m^*(f) < M_m^*(f_m^*).$$

For any $f \in \mathcal{F}$ with $D(f_m^* \| f) \geq \varepsilon$, we have

$$\begin{aligned} M_m^*(f_m^*) &= \mathbb{E}_{\bar{f}} \log \bar{f}(Z) \\ &= D(\bar{f} \| f) - D(\bar{f} \| f) + \mathbb{E}_{\bar{f}} \log \bar{f}(Z) \\ &\geq \varepsilon + M_m^*(f), \end{aligned}$$

since $D(\bar{f} \| f) = D(f_m^* \| f) \geq \varepsilon$. \square

Alternative proof of Theorem 5.1. Suppose without loss of generality that H_m is true. Then by exchangeability, the bFDR of the SL method is

$$\mathbb{P}(H_{(R_\alpha)} \text{ is true}, R_\alpha > 0) = m\bar{\pi}_0 \mathbb{P}(p_{(R_\alpha)} = p_m, R_\alpha > 0).$$

Let $q_{(1)} \leq \dots \leq q_{(m-1)}$ denote the order statistics of p_1, \dots, p_{m-1} , and note that p_m achieves the maximum in (12) as the $(k+1)$ th order statistic if $q_{(k)} < p_m < q_{(k+1)}$ and

$$\frac{\alpha(k+1)}{m} - p_m > \left[\max_{j=k+1, \dots, m-1} \left\{ \Delta_j + \frac{\alpha}{m} \right\} \right] \vee \left[\max_{j=0, \dots, k} \Delta_j \right],$$

for $k \leq m-1$, where $q_{(0)} := 0$ and $\Delta_j := \frac{\alpha j}{m} - q_{(j)}$. Rearranging the above inequalities gives the range in which p_m achieves the maximum and is equal to the $(k+1)$ th order statistic, i.e. $q_{(k)} < p_m$ and

$$p_m < \frac{\alpha k}{m} - \left[\max_{j=k+1, \dots, m-1} \Delta_j \right] \vee \left[\max_{j=0, \dots, k} \Delta_j - \frac{\alpha}{m} \right].$$

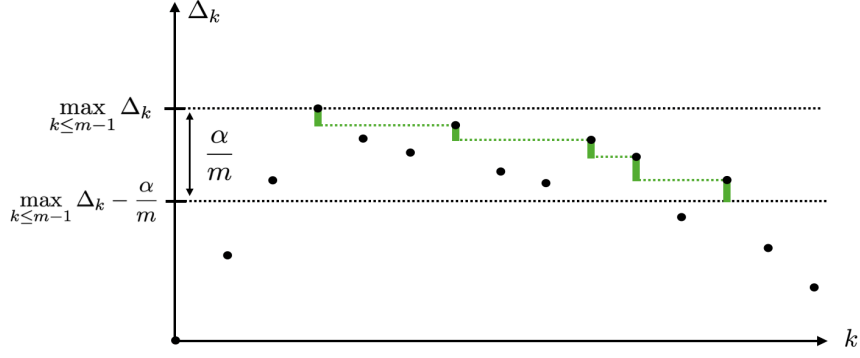


Figure 10: Each length of the interval range in which p_m achieves the maximum in (12) is indicated by a vertical green bar, and the sum of these lengths is $\frac{\alpha}{m}$.

This range is non-empty when Δ_k exceeds each of $\Delta_{k+1}, \dots, \Delta_{m-1}$ as well as $\max_{j=0, \dots, m-1} \Delta_j - \frac{\alpha}{m}$, and has length

$$\Delta_k - \left[\max_{j=k+1, \dots, m-1} \Delta_j \right] \vee \left[\max_{j=0, \dots, k} \Delta_j - \frac{\alpha}{m} \right]$$

The sum of lengths of the non-empty ranges is telescoping and equal to $\frac{\alpha}{m}$, as illustrated in Figure 10. \square

Proof of Proposition 5.1. When $Z \sim g_\theta$, the density of $p = 1 - G_{\theta_0}(Z)$ is

$$\frac{d}{dt} \mathbb{P}_\theta(p \leq t) = \frac{g_\theta}{g_{\theta_0}}(G_{\theta_0}^{-1}(1-t)).$$

At $\theta = \theta_0$, the above ratio is equal to 1. When $\theta \leq \theta_0$, the log density has a positive derivative in θ when

$$\frac{d}{d\theta} \left[\log \frac{g_\theta}{g_{\theta_0}}(G_{\theta_0}^{-1}(1-t)) \right] = G_{\theta_0}^{-1}(1-t) - \mathbb{E}_\theta(Z) > 0$$

which holds for all $t \leq \alpha$ if $G_{\theta_0}^{-1}(1-\alpha) > \mathbb{E}_{\theta_0}(Z)$. \square

Proposition B.1. *In the setting of section A.1, the best PE decision rule for minimizing the weighted classification risk is defined by (16) and (17).*

Proof. For any PE decision rule δ ,

$$\begin{aligned} \tilde{\mathbb{E}}L_\lambda(\tilde{H}, \delta(\tilde{\mathbf{p}})) &= \frac{1}{m!} \sum_{\sigma \in \mathcal{S}_m} \mathbb{E}L_\lambda(H_\sigma, \delta(\mathbf{p}_\sigma)) \\ &= \frac{1}{m!} \sum_{\sigma \in \mathcal{S}_m} \mathbb{E}L_\lambda(H_\sigma, \delta(\mathbf{p})_\sigma) \\ &= \mathbb{E}L_\lambda(H, \delta(\mathbf{p})). \end{aligned}$$

The Bayes rule

$$\delta^* = \operatorname{argmin}_{\delta} \mathbb{E}[L_{\lambda}(\tilde{H}, \delta(\tilde{\mathbf{p}})) \mid \tilde{\mathbf{p}}],$$

is itself PE due to exchangeability of $(\tilde{H}_i, \tilde{\mathbf{p}}_i)$ across $i = 1, \dots, m$. To see this, note that for any $\sigma \in \mathcal{S}_m$, we have

$$(\tilde{H}, \tilde{\mathbf{p}}) \stackrel{(d)}{=} (\tilde{H}_{\sigma}, \tilde{\mathbf{p}}_{\sigma}),$$

which implies the posterior probability mass function of $\tilde{H} \mid \tilde{\mathbf{p}} = t$ at $\tilde{h} \in \{0, 1\}^m$ is equal to the posterior pmf of $\tilde{H}_{\sigma} \mid \tilde{\mathbf{p}}_{\sigma} = t_{\sigma}$ at \tilde{h}_{σ} . Thus,

$$\begin{aligned} \delta^*(t) &= \operatorname{argmin}_{h \in \{0, 1\}^m} \mathbb{E}[L_{\lambda}(\tilde{H}_{\sigma}, h) \mid \tilde{\mathbf{p}}_{\sigma} = t] \\ &= \operatorname{argmin}_{h \in \{0, 1\}^m} \mathbb{E}[L_{\lambda}(\tilde{H}, h_{\sigma^{-1}}) \mid \tilde{\mathbf{p}}_{\sigma} = t] \\ &= \left[\operatorname{argmin}_{g \in \{0, 1\}^m} \mathbb{E}[L_{\lambda}(\tilde{H}, g) \mid \tilde{\mathbf{p}} = t_{\sigma^{-1}}] \right]_{\sigma} \\ &= \delta^*(t_{\sigma^{-1}})_{\sigma}. \end{aligned}$$

Since the above holds for any permutation σ , the Bayes rule in model (15) is a PE decision rule. Since the average risk in the Bayes model (15) is equal to the risk function in the frequentist model for every configuration of truth values $H \in \{0, 1\}^m$, the Bayes rule is equal to the best PE rule

$$\delta^* = \operatorname{argmin}_{\delta \text{ PE}} \mathbb{E}L_{\lambda}(H, \delta(\mathbf{p})). \quad \square$$

Proof of Proposition A.1. According to Bayes rule, $\mathbb{P}(\theta_i = 0 \mid p_i = t, G_m)$ is equal to

$$= \frac{\mathbb{P}(\theta_i = 0 \mid G_m) f_0(t)}{\sum_{k=1}^m f_{\theta_{(k)}}(t) \mathbb{P}(\operatorname{rank}(\theta_i) = k \mid G_m)},$$

where $\operatorname{rank}(\theta_i) = k$ when $\theta_j < \theta_i$ exactly $k - 1$ indices $j \in [m]$, and $\theta_{(1)} \leq \dots \leq \theta_{(m)}$ are the ordered values of $\theta_1, \dots, \theta_m$. Since $\theta_1, \dots, \theta_m$ are exchangeable, the above is equal to

$$\mathbb{P}(\theta_i = 0 \mid p_i = t, G_m) = \frac{G_m(\{0\}) f_0(t)}{\frac{1}{m} \sum_{j=1}^m f_{\theta_j}(t)}.$$

For (20), note that when $p_i = p_{(k)}$, exchangeability implies

$$\begin{aligned} &\mathbb{P}(\theta_i = 0 \mid p_1 = t_1, \dots, p_m = t_m, G_m) \\ &\propto \sum_{\pi \in \mathcal{S}_m: \theta_{\pi(i)} = 0} \prod_{j=1}^m f_{\theta_{\pi(j)}}(t_j). \end{aligned}$$

□

Proof of Lemma A.1. The argument is adapted from Theorem 3.1 in [Greenshtein and Ritov \(2009\)](#). Supposing without loss of generality that $H_1 = 0$ and $H_2 = 1$,

$$\text{clfd}_i(\mathbf{p}) = \frac{\bar{\pi}_0 f_0(p_i)}{\bar{\pi}_0 f_0(p_i) + \bar{\pi}_1 f_1(p_i) \cdot X_i},$$

where X_i is a likelihood ratio,

$$X_i := \frac{\sum_{\sigma \in \mathcal{S}_m: \sigma(i)=2} \prod_{j \in [m] \setminus \{i\}}^m f^{(\sigma(j))}(p_j)}{\sum_{\sigma \in \mathcal{S}_m: \sigma(i)=1} \prod_{j \in [m] \setminus \{i\}}^m f^{(\sigma(j))}(p_j)}$$

for testing between the following two hypotheses:

Hyp₀ : Observe a random permutation of p_{-i} when $H_i = 1$

Hyp₁ : Observe a random permutation of p_{-i} when $H_i = 0$,

where the permutations are drawn uniformly at random from \mathcal{S}_{m-1} . A simpler testing problem is:

$$\begin{aligned} \widetilde{\text{Hyp}}_0 &: \tilde{p}_1, \dots, \tilde{p}_{m_0} \stackrel{\text{iid}}{\sim} f_0, \text{ and } (\tilde{p}_{m_0+1}, \dots, \tilde{p}_{m-1}) \stackrel{\text{iid}}{\sim} f_1 \\ \widetilde{\text{Hyp}}_1 &: \frac{1}{m_0} \sum_{\ell=1}^{m_0} \left[(\tilde{p}_{1:m_0})_{-\ell} \stackrel{\text{iid}}{\sim} f_0, (\tilde{p}_\ell, \tilde{p}_{m_0+1}, \dots, \tilde{p}_{m-1}) \stackrel{\text{iid}}{\sim} f_1 \right], \end{aligned}$$

since Hyp₀, Hyp₁ can be obtained from $\widetilde{\text{Hyp}}_0, \widetilde{\text{Hyp}}_1$ by adding a random permutation. If $H_i = 0$ (resp. $H_i = 1$), then the distribution of X_i is as if the data were generated by Hyp₁ (resp. Hyp₀). The likelihood ratio of $\widetilde{\text{Hyp}}_1$ to $\widetilde{\text{Hyp}}_0$ has variance

$$\text{Var}_0 \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \frac{f_1}{f_0}(\tilde{p}_\ell) \right) = \frac{1}{m_0} \text{Var}_0 \left(\frac{f_1}{f_0}(p_1) \right) \rightarrow 0$$

by assumption, where Var_0 denotes the variance operation when $\widetilde{\text{Hyp}}_0$ holds. It follows from Lemma 2.1 in [Greenshtein and Ritov \(2009\)](#) that

$$\mathbb{E}_{H_i=1}(X_i - 1)^2 \leq \widetilde{\mathbb{E}}_0 \left(\frac{1}{m_0} \sum_{\ell=1}^{m_0} \frac{f_1}{f_0}(\tilde{p}_\ell) - 1 \right)^2 \rightarrow 0.$$

A symmetric argument yields

$$\mathbb{E}_{H_i=0}(X_i - 1)^2 \leq \widetilde{\mathbb{E}}_1 \left(\frac{1}{m_1} \sum_{\ell=m_0}^{m-1} \frac{f_0}{f_1}(\tilde{p}_\ell) - 1 \right)^2 \rightarrow 0,$$

under the condition that $\text{Var} \left(\frac{f_0}{f_1}(p_2) \right) < \infty$ when $H_2 = 1$. Here we are abusing notation by writing the index ℓ from m_0 to $m-1$, to denote summing over the $m_1 - 1$ many p -values

drawn from f_1 in the scenario described by $\widetilde{\text{Hyp}}_1$. It now follows from Chebyshev's inequality that

$$\begin{aligned}
& \mathbb{P} \left(\left| \frac{\text{clfdr}_i(\mathbf{p})}{\text{lfdr}(p_i)} - 1 \right| > m^{-1/2} (\log m)^{3/2} \right) \\
& \leq \mathbb{P} \left(|X_i - 1| > m^{-1/2} (\log m)^{3/2} \right) \\
& \leq \frac{m}{(\log m)^3} \cdot \frac{\text{Var} \left(\frac{f_0}{f_1}(p_2) \right) \vee \text{Var} \left(\frac{f_1}{f_0}(p_1) \right)}{m_0 \wedge m_1} \\
& \leq \frac{C}{(\log m)^3}
\end{aligned}$$

for some constant $C > 0$ as $m \rightarrow \infty$, since $\bar{\pi}_0$ is bounded away from zero and one. \square

Proof of Lemma A.2. By the law of total probability,

$$\begin{aligned}
\mathbb{P}(\text{rank}(p_m) = R) &= \sum_{\ell=1}^L \frac{1}{L} \cdot \mathbb{P}(\text{rank}(p_m) = R \mid p_m = \ell/L) \\
&= \sum_{\ell=1}^L \frac{1}{L} \cdot 1_{\{\hat{\tau}_\alpha(\ell/L) = \ell/L\}} \cdot \frac{1}{n_\ell + 1},
\end{aligned}$$

where $n_\ell := \#\{i < m : p_i = \ell/L\}$, and we have used explicit notation $\hat{\tau}_\alpha(p_m)$ to denote the threshold $\hat{\tau}_\alpha$ as a function of p_m ,

$$\begin{aligned}
\hat{\tau}_\alpha(p_m) &:= \operatorname{argmax}_{p(k)} \left\{ \frac{\alpha k}{m} - p(k) \right\} \\
&= \frac{1}{L} \cdot \operatorname{argmax}_{\ell=0, \dots, L} \left\{ \frac{\alpha L}{m} \cdot \#\{i \leq m : p_i \leq \ell/L\} - \ell \right\},
\end{aligned}$$

treating p_1, \dots, p_{m-1} as non-random elements of the grid $\{1/L, \dots, L/L\}$. Define

$$\Delta_\ell := \frac{\alpha L}{m} N_\ell - \ell, \quad \ell = 1, \dots, L$$

where $N_\ell := \#\{i < m : p_i \leq \ell/L\}$, and let $\ell^* := \operatorname{argmax}_\ell \Delta_\ell$. We claim that

$$\hat{\tau}_\alpha(\ell/L) = \ell/L \iff \Delta_\ell > \left[\Delta_{\ell^*} - \frac{\alpha L}{m} \right] \vee \max_{k > \ell} \Delta_k, \quad (26)$$

which follows from the same argument as in the alternative proof of Theorem 5.1 in Appendix B. The claim implies

$$\begin{aligned}
\mathbb{P}(\text{rank}(p_m) = R) &= \sum_{\ell=1}^L \frac{1}{L} \cdot 1_{\{\hat{\tau}_\alpha(\ell/L) = \ell/L\}} \cdot \frac{1}{n_\ell + 1} \\
&\leq \frac{1}{L} \sum_{\ell=1}^L 1_{\{\Delta_\ell > [\Delta_{\ell^*} - \frac{\alpha L}{m}] \vee \max_{k > \ell} \Delta_k\}}.
\end{aligned}$$

Without loss of generality, suppose that $p_1 \leq p_2 \leq \dots \leq p_{m-1}$. If $p_i < p_j$ are among these p -values and both $\Delta_{p_i L}, \Delta_{p_j L}$ satisfy the rhs of (26), then as $L \rightarrow \infty$, it follows from the definition of Δ_ℓ that as L grows,

$$\Delta_\ell - \Delta_{\ell+1} = 1 \quad \text{for all but a fixed number of } \ell \in (Lp_i, Lp_j).$$

Thus for the distinct values $\ell^* =: \ell_1 < \ell_2 < \dots < \ell_k$ for which the rhs of (26) holds, we must have $\Delta_{\ell_i} - \Delta_{\ell_{i+1}} = 1$ for all but a fixed number of winners ℓ_i , the number of which is less than m . Therefore

$$\begin{aligned} \mathbb{P}(\text{rank}(p_m) = R) &\leq \frac{1}{L} \sum_{\ell=1}^L 1_{\{\Delta_\ell > [\Delta_{\ell^*} - \frac{\alpha L}{m}] \vee \max_{k > \ell} \Delta_k\}} \\ &\leq \frac{1}{L} \left(m + \sum_{i=1}^{k-1} (\Delta_i - \Delta_{i+1}) + \Delta_k - \Delta_{\ell^*} + \frac{\alpha L}{m} \right) \\ &= \frac{1}{L} \left(m + \frac{\alpha L}{m} \right) = \frac{\alpha}{m} + O\left(\frac{m}{L}\right), \end{aligned}$$

as $L \rightarrow \infty$. □

Proof of Theorem A.2. By symmetry of the nulls,

$$\mathbb{P}(H_{(R_\alpha)} = 0) = m_0 \mathbb{P}(\text{rank}(p_m) = R).$$

By the argument in Theorem A.1, the probability that p_m is the R^{th} smallest p -value (breaking ties uniformly at random) is

$$\mathbb{P}(\text{rank}(p_m) = R) = \mathbb{E} \left(\frac{1}{L} \sum_{\ell=1}^L 1_{\{\Delta_\ell > [\Delta_{\ell_m^*} - \alpha L/m] \vee \max_{j > \ell} \Delta_j\}} \cdot \frac{1}{n_\ell + 1} \right),$$

where $n_\ell := \#\{i < m : p_i = \ell/L\}$ and Δ_ℓ are defined

$$\Delta_\ell := \frac{\alpha L}{m} N_\ell - \ell, \quad \ell = 0, \dots, L,$$

with $N_\ell := \sum_{k=1}^\ell n_k$ and $\ell_m^* := \arg\max_{\ell=0, \dots, L} \Delta_\ell$. As $m \rightarrow \infty$, we have the following convergence in probability,

$$\frac{n_\ell}{m f^*(\ell/L)} \xrightarrow{p} 1, \quad \Delta_\ell \xrightarrow{p} \alpha L \sum_{k=0}^\ell f^*(k/L) - \ell.$$

Since the maximizer of $\alpha L \sum_{k=0}^\ell f^*(k/L) - \ell$ is unique, we have $\ell_m^* = \ell^*$ with high probability as $m \rightarrow \infty$, and

$$m f^*(\ell^*/L) \cdot \sum_{\ell=1}^L 1_{\{\Delta_\ell > [\Delta_{\ell_m^*} - \alpha L/m] \vee \max_{j > \ell} \Delta_j\}} \cdot \frac{1}{n_\ell + 1} \xrightarrow{\mathbb{P}} 1_{\{\ell^* > 0\}},$$

from which it follows that

$$\mathbb{E} \left[\frac{m_0}{mf^*(\ell^*/L)} \cdot \frac{L}{L} \cdot mf^*(\ell^*/L) \cdot \sum_{\ell=1}^L 1_{\{\Delta_\ell > [\Delta_{\ell_m^*} - \alpha L/m] \vee \max_{j>\ell} \Delta_j\}} \cdot \frac{1}{n_\ell + 1} \right] \rightarrow \frac{\pi_0^*}{Lf^*(\ell^*/L)},$$

when $\ell^* > 0$, and zero otherwise. \square

B.0.1 Counterexample to the discrete-uniform assumption

The SL guarantee breaks down in the setting where each null p -value is uniformly distributed on the grid $\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{L-1}{L}, 1\}$.

Counterexample. Let $m = 6, L = 9, \alpha = 1/2$, and the alternative p -values are

$$\begin{aligned} p_1 &= p_2 = 1/L \\ p_i &= i/L \quad i = 2, 3, 4. \end{aligned}$$

Then the probability that $p_m \sim \text{Uniform}\{\frac{1}{L}, \frac{2}{L}, \dots, \frac{L-1}{L}, 1\}$ is the last $\text{SL}(\alpha)$ rejection is

$$\begin{aligned} \mathbb{P}(p_{(R_\alpha)} = p_m) &= \sum_{\ell=1}^L \frac{1}{L} \mathbb{P}(p_{(R_\alpha)} = p_m \mid p_m = \ell/L) \\ &= \sum_{\ell=1}^L \frac{1}{L} \cdot 1_{\{p_{(R_\alpha)} = \ell/L\}} \cdot \frac{1}{n_\ell + 1}, \end{aligned}$$

where $n_\ell := \#\{i < m : p_i = \ell/L\}$ for $\ell = 1, \dots, L$. It is straightforward to check that $1_{\{p_{(R_\alpha)} = \ell/L\}} = 1$ for $\ell = 1, 2, 3, 4$, so that the above evaluates to

$$\mathbb{P}(p_{(R_\alpha)} = p_m) = \frac{1}{9} \left(\frac{1}{3} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \right) = 0.204 > 0.167 = \frac{2\alpha}{m}.$$

\square

B.0.2 Counterexample to the super-uniform assumption

The SL guarantee may break down when the nulls are super-uniformly distributed instead of $\text{Uniform}(0, 1)$ distributed, as illustrated by the following example.

Counterexample. Let $m = 2, H_1 = 0$, and $H_2 = 1$, with $p_1 \sim f^{(1)}$ defined

$$f^{(1)}(t) := \begin{cases} \frac{1}{2} & 0 \leq t \leq \frac{1}{4} \\ \frac{3}{2} & \frac{1}{4} < t \leq \frac{1}{2} \\ 1 & \frac{1}{2} < t \leq 1, \end{cases}$$

and $p_2 \equiv \frac{1}{4}$. It is straightforward to verify that $f^{(1)}$ is super-uniform, i.e.

$$\int_0^t f^{(1)}(s) ds \leq t \quad \text{for any } t \in [0, 1].$$

The event $p_{(R_\alpha)} = p_1$ is equivalent to

$$\left(\{p_1 \leq p_2\} \cap \left\{p_1 - \frac{\alpha}{2} < (p_2 - \alpha) \wedge 0\right\}\right) \cup \left(\{p_1 > p_2\} \cap \left\{p_1 - \alpha < \left(p_2 - \frac{\alpha}{2}\right) \wedge 0\right\}\right).$$

Plugging in $\alpha = \frac{1}{2}$ and $p_2 = \frac{1}{4}$ gives

$$H_{(R_\alpha)} = 0 \iff p_{(R_\alpha)} = p_1 \iff p_1 \in (1/4, 1/2),$$

which occurs with probability $\frac{3}{2} \times \frac{1}{4} = \frac{3}{8} > \frac{\alpha}{2} = \frac{1}{4}$. □

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Roy. Statist. Soc. Ser. B* **57**(1): 289–300.
URL: [http://links.jstor.org/sici?sici=0035-9246\(1995\)57:1;289:CTFDRA;2.0.CO;2-Eorigin=MSN](http://links.jstor.org/sici?sici=0035-9246(1995)57:1;289:CTFDRA;2.0.CO;2-Eorigin=MSN)
- Dawid, A. P. (1982). The well-calibrated Bayesian, *J. Amer. Statist. Assoc.* **77**(379): 605–613.
URL: [http://links.jstor.org/sici?sici=0162-1459\(198209\)77:379;605:TWB;2.0.CO;2-Qorigin=MSN](http://links.jstor.org/sici?sici=0162-1459(198209)77:379;605:TWB;2.0.CO;2-Qorigin=MSN)
- Eaton, M. L. and George, E. I. (2021). Charles stein and invariance: Beginning with the hunt–stein theorem, *The Annals of Statistics* **49**(4): 1815–1822.
- Efron, B. (2019). Bayes, oracle Bayes and empirical Bayes, *Statist. Sci.* **34**(2): 177–201.
- Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *J. Amer. Statist. Assoc.* **96**(456): 1151–1160.
- Efron, B., Turnbull, B. B. and Narasimhan, B. (2011). locfdr: Computes local false discovery rates, *R package version 1*: 1–7.
- Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control, *Ann. Statist.* **32**(3): 1035–1061.
- Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy, *Ann. Intern. Med.* **130**(12): 995–1004.
- Greenshtein, E. and Ritov, Y. (2009). Asymptotic efficiency of simple decisions for the compound decision problem, *Lecture Notes-Monograph Series* pp. 266–275.
- Grenander, U. (1956). On the theory of mortality measurement: Part II, *Scandinavian Actuarial Journal* **1956**(2): 125–153.

- Gupta, C., Podkopaev, A. and Ramdas, A. (2020). Distribution-free binary classification: prediction sets, confidence intervals and calibration, *Adv. Neural Inf. Process. Syst.* **33**: 3711–3723.
- Han, Y. and Niles-Weed, J. (2024). Approximate independence of permutation mixtures, *arXiv preprint arXiv:2408.09341*.
- Hannan, J. F. and Robbins, H. (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions, *Ann. Math. Statist.* **26**: 37–51.
- Hung, K. and Fithian, W. (2020). Statistical methods for replicability assessment, *Ann. Appl. Stat.* **14**(3): 1063–1087.
URL: <https://doi.org/10.1214/20-AOAS1336>
- Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O. A. and Heller, R. (2022). Quantifying replicability and consistency in systematic reviews, *Stat. Biopharm. Res.* pp. 1–14.
- Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical bayes estimation of normal means, *The Annals of Statistics* **37**(4): 1647–1684.
- Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters, *The Annals of Mathematical Statistics* pp. 887–906.
- Lindsey, J. (1974). Construction and comparison of statistical models, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(3): 418–425.
- Maier, M., Bartoš, F., Stanley, T., Shanks, D. R., Harris, A. J. and Wagenmakers, E.-J. (2022). No evidence for nudging after adjusting for publication bias, *Proc. Natl. Acad. Sci. U.S.A.* **119**(31): e2200300119.
- McCullagh, P. (2014). An asymptotic approximation for the permanent of a doubly stochastic matrix, *Journal of Statistical Computation and Simulation* **84**(2): 404–414.
- Mertens, S., Herberz, M., Hahnel, U. J. and Brosch, T. (2022a). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains, *Proc. Natl. Acad. Sci. U.S.A.* **119**(1): e2107346118.
- Mertens, S., Herberz, M., Hahnel, U. J. and Brosch, T. (2022b). Reply to Maier et al., Szaszi et al., and Bakdash and Marusich: The present and future of choice architecture research, *Proc. Natl. Acad. Sci. U.S.A.* **119**(31): e2202928119.
- Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems, *Proc. of the Second Berkeley Sympos. on Math. Statist. and Probability*, Univ. California Press, Berkeley-Los Angeles, Calif., pp. 131–148.
- Savage, L. J. (1972). *The foundations of statistics*, Courier Corp.

- Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*, SIAM.
- Soloff, J. A., Xiang, D. and Fithian, W. (2024). The edge of discovery: Controlling the local false discovery rate at the margin, *Ann. Statist.* **52**(2): 580–601.
- Song, S. J., Amir, A., Metcalf, J. L., Amato, K. R., Xu, Z. Z., Humphrey, G. and Knight, R. (2016). Preservation methods differ in fecal microbiome stability, affecting suitability for field studies, *mSystems* **1**(3): 10.1128/msystems.00021–16.
- Storey, J. D. (2002). A direct approach to false discovery rates, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**(3): 479–498.
URL: <https://doi.org/10.1111/1467-9868.00346>
- Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value, *The annals of statistics* **31**(6): 2013–2035.
- Storey, J. D., Taylor, J. E. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66**(1): 187–205.
- Strimmer, K. (2008). A unified approach to false discovery rate estimation, *BMC bioinformatics* **9**(1): 1–14.
- Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control, *J. Amer. Statist. Assoc.* **102**(479): 901–912.
URL: <https://doi.org/10.1198/016214507000000545>
- Szaszi, B., Higney, A., Charlton, A., Gelman, A., Ziano, I., Aczel, B., Goldstein, D. G., Yeager, D. S. and Tipton, E. (2022). No reason to expect large and consistent effects of nudge interventions, *Proc. Natl. Acad. Sci. U.S.A.* **119**(31): e2200732119.
- Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*, Penguin.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*, Vol. 3, Cambridge university press.
- Weinstein, A. (2021). On permutation invariant problems in large-scale inference, *arXiv preprint arXiv:2110.06250*.
- Weinstein, A., Wallin, J., Yekutieli, D. and Bogdan, M. (2019). Nonparametric shrinkage estimation in high dimensional generalized linear models via polya trees, *arXiv e-prints* pp. arXiv–1908.
- Zhang, C.-H. (2003). Compound decision theory and empirical Bayes methods, *Ann. Statist.* **31**(2): 379–390.
URL: <https://doi.org/10.1214/aos/1051027872>

Zhang, C.-H. (2009). Generalized maximum likelihood estimation of normal mixture densities, *Statistica Sinica* pp. 1297–1318.