# A frequentist local false discovery rate

Daniel Xiang[1], Jake A. Soloff[1], and William Fithian[2]

[1]*Department of Statistics, University of Chicago*
[2]*Department of Statistics, University of California, Berkeley*

September 29, 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 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 yields a calibrated forecast of whether its null hypothesis is true. Our definition arises naturally from compound decision theory, yielding the best separable decision rule under the weighted classification loss, and it can be estimated efficiently in finite samples using parametric or non-parametric methods. 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 concrete scientific hypothesis is a random variable that rises and falls according to an observer's prejudices (Goodman, 1999; Savage, 1972).

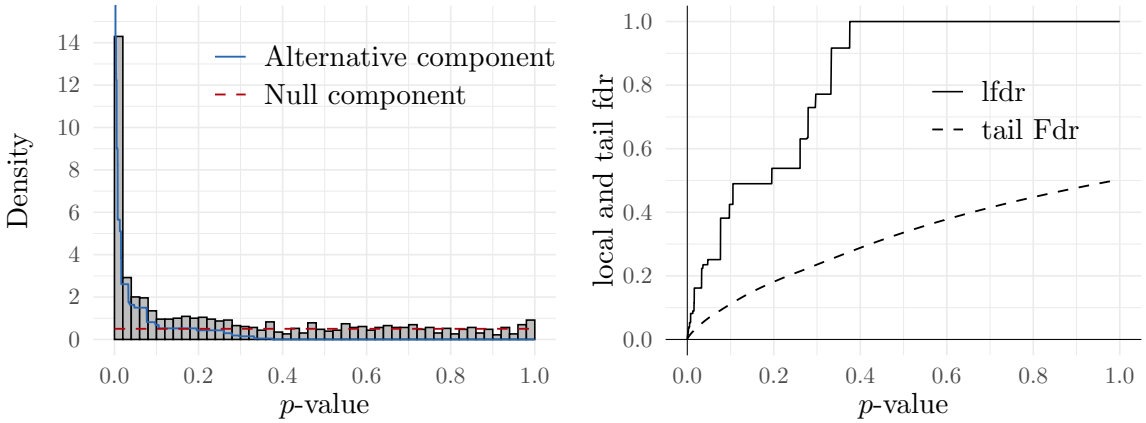


Figure 1: Microbiome preservation example. The left panel shows the histogram for $m = 1147$ stratified permutation test $p$-values comparing relative abundance in fresh vs. eight-week-old samples, along with a nonparametric estimate of the null and alternative components of the mixture density via the empirical Bayes estimator of Strimmer (2008). The right panel shows the corresponding estimates of the local false discovery rate and tail false discovery rate. Due to the taxonomic structure in the data and the correlations between different species, it would be difficult to justify these lfdr estimates in a Bayesian analysis, but they have natural interpretations in our frequentist framework.

Both frequentists and Bayesians are better positioned to answer the question when the hypothesis is one of many, as long as the other hypotheses are *relevant*: that is, the cases are sufficiently alike to justify a combined analysis. Then, hierarchical or empirical Bayesian methods are appealing because they effectively replace the analyst's prior with one that is learned from the data. However, these methods still require the analyst to model the truth status of individual hypotheses as random variables, and to mathematically formalize the assumption of relevance, typically by assuming either that the hypotheses (as well as the test statistics) are exchangeable, or by introducing a parametric model for their dependence. In a concrete scientific context where the hypotheses are not perfectly alike *a priori*, these may be difficult assumptions for the analyst to accept.

Frequentists can circumvent the need for priors by controlling the false discovery rate (FDR), which measures the expected fraction of true nulls among hypotheses with test statistics falling into the rejection region of a multiple testing procedure (Benjamini and Hochberg, 1995). The $q$-values of Storey (2002) even give a kind of FDR estimate for individual hypotheses. However, instead of answering our original question, $q$-values answer a substantively different one, roughly: among all hypotheses with test statistics as extreme *or more extreme* than this one, what fraction are true nulls? As we discuss in Section 6, these are two very different questions. Misinterpreting the $q$-value as a measure of confidence in a given discovery would make us systematically, and often severely, over-optimistic.

In this work we propose a new answer to our motivating question that blends the Bayesian

and frequentist approaches. Suppose that we observe $m$ test statistics $z_1, \ldots, z_m$ for null hypotheses $H_1, \ldots, H_m$, of which $m_0$ are true nulls. For $i = 1, \ldots, m$, let $f^{(i)}(t)$ denote the density of $z_i$, and define the (frequentist) *local false discovery rate* (lfdr) as the relative frequency of null test statistics at each point in the sample space:

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

If $m$ is large, this ratio is approximately the proportion of nulls among all hypotheses whose test statistics fall in a small neighborhood around $t$.

In the common case where the null test statistics share the same density $f_0$, we have the simpler expression

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

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

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

is the average density.

The null density $f_0$ could, for example, represent the standard Gaussian distribution on the real line $\mathbb{R}$, if the statistics are $z$-values, or the uniform distribution on the unit interval $[0, 1]$ if they are $p$-values. Let $\mathcal{Z}$ generically denote the common sample space where $z_1, \ldots, z_m$ are realized. For simplicity of exposition, we will assume throughout that the test statistics are continuous, but most of our results extend naturally to discrete sample spaces.

Readers familiar with Efron et al. (2001) will note that the expression (2) appears nearly identical to the definition of lfdr in the Bayesian two-groups model (which we review in Section 3.1), but there are key differences. Perhaps most importantly, our lfdr is defined without reference to any Bayesian prior; it depends only on the marginal densities of the $m$ statistics. As such, our definition of $\text{lfdr}(t)$ does not represent a Bayesian posterior probability that $H_i$ is true given $z_i = t$. In our frequentist model, that probability is always either zero or one.

However, the frequentist lfdr does answer our motivating question, in a sense: it represents the conditional probability, given that *some* test statistic equals $t$, that the corresponding null hypothesis is true:

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

If the conditional probability in (3) at first appears paradoxical in our frequentist setting, note that the truth status of $H_J$ is random because the index $J$ is random. We prove the relation (3) in Section 3, where we compare our lfdr with the Bayesian lfdr in more detail.

In Section 3.2, we establish two other appealing properties of the lfdr that reinforce its conceptual usefulness: First, we show that $\text{lfdr}(z_1), \ldots, \text{lfdr}(z_m)$ are calibrated "forecasts" for the truth of $H_1, \ldots, H_m$. The lfdr function gives the sharpest forecasts of any calibrated transformation from the sample space $\mathcal{Z}$ to the unit interval.

Second, we show that our lfdr arises naturally in compound decision theory for multiple testing: if a Type I error is $\lambda$ times as costly as a Type II error, the best separable rejection rule (the best rule where we decide whether to reject $H_i$ using only $z_i$) rejects $H_i$ if and only if $\text{lfdr}(z_i) \leq 1/(1+\lambda)$ — coinciding with the optimal decision rule in a Bayesian model where $\text{lfdr}(z_i)$ represents the Bayesian posterior probability that $H_i$ is true.

Although the frequentist lfdr depends on unknown quantities, it can usually be estimated efficiently from the data if $m$ is reasonably large and the dependence between test statistics is not too strong. 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 5 discusses approaches to this problem based on standard parametric or nonparametric methods for density estimation in the i.i.d. setting, and argues classical empirical Bayes methods commonly applied under the two-groups model can be understood as estimates of the frequentist lfdr. Section 6 shows that the support line procedure of Soloff et al. (2024), which is closely related to Strimmer's monotone lfdr estimator Strimmer (2008), succeeds in finite samples at controlling the *boundary FDR*, an error criterion closely related to the lfdr, provided the statistics are independent $p$-values.

The next section discusses two motivating examples in which the frequentist lfdr is a useful concept to have available.

## 2 Motivating examples

### 2.1 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 $\Omega$ 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 $i^{\text{th}}$ and $j^{\text{th}}$ 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 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.

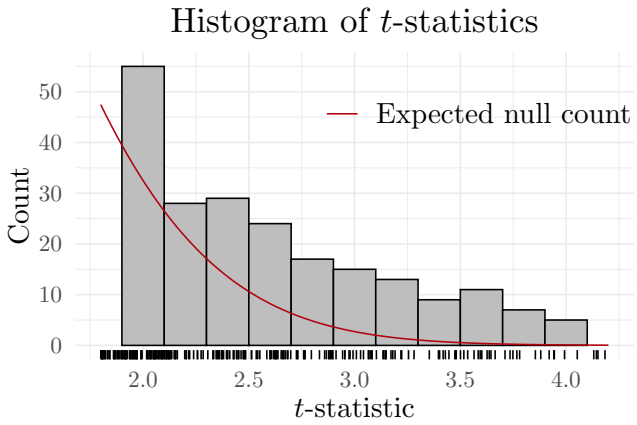


Figure 2: Histogram of $t$-statistics in the GGM example, with dimensions: $d = 80$, $n = 10d$, $m = 3,160$. The null distribution is $t_{n-d}$.

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 2 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%. Without a Bayesian prior on the entries of $\Omega$, we are somehow still able to quantify our confidence in a typical null hypothesis whose $t$-statistic is near 3. We discuss several interpretations in Section 3.

In the next section, we analyze a real dataset from biology to further illustrate interesting distinctions between our approach and others.

## 2.2 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 species[1], 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.

# 3 Interpreting the frequentist lfdr

## 3.1 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 $\pi_0$ 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:

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

[1]We 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.

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

When $\pi_0$, $f_0$, and $f_1$ are fixed and known, the posterior probabilities $\text{lfdr}^*(z_1), \ldots, \text{lfdr}^*(z_m)$ fully describe the posterior, and they represent the sharpest calibrated forecasts for the truth status of the hypotheses $H_1, \ldots, H_m$ (Dawid, 1982; Gupta et al., 2020). They also have a natural interpretation in decision theory. For a decision rule $\delta$ that returns an accept/reject decision for each hypothesis, define the *weighted classification loss*, which penalizes the analyst $\lambda$ 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 $\delta$ rejects a true $H_i$, and a false negative occurs when $\delta$ 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 $\pi_0$ are unknown. If $\pi_0 \approx 1$, then estimating the marginal density $f(t)$ from the i.i.d. sample $z_1, \ldots, z_m$ 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 $H_1, \ldots, H_m$ 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 $\lambda$ 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.

### 3.2 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 3.1.** *Suppose $z_1, \ldots, z_m$ 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 $\text{lfdr}(t)$. 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 4 for further discussion of connections between the lfdr, FDR, and mFDR.

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

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

**Theorem 3.2.** *Let $\ell_i \coloneqq \text{lfdr}(z_i)$ for $i = 1, \ldots, m$ and suppose $z_1, \ldots, z_m$ 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 $g : \mathbb{R} \to [0, 1]$ is calibrated, then for any $t$,*

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

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

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

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

**Theorem 3.3.** *If $\delta$ is a separable decision rule, i.e. $\delta_i(z_1, \ldots, z_m) = g(z_i)$ for some univariate function $g$, then*

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

*where*

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

### 3.3 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, especially when the test-statistics are dependent. We discuss the problem of estimating the lfdr in section 5, 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 doesn't account for additional information known by the analyst. The conditional probability interpretation only applies to an analyst who doesn't know which null hypothesis corresponds to the test-statistic realized at, e.g. $z_J = 3$. The analyst who does know which $H_J$ has $z_J = 3$ may arrive at a different conclusion, particularly if their prior is non-exchangeable. A large difference between these two conclusions suggests that we might be better off analyzing the hypotheses in sub-groups, conditional on some covariate.

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 $\{z_1, \ldots, z_m\}$ 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.

## 4 lfdr and FDR

We begin this section by continuing the numerical experiment in the Gaussian graphical model from Section 2.1. This example illustrates the bias of $q$-values as a measure of confidence over a wide range of the sample space.

Recall the precision matrix $\Omega$ that defines our null hypotheses: $H_{ij}$ is true if there is zero partial correlation between $X_i$ and $X_j$, i.e. $\Omega_{ij} = 0$. From the list of $t$-statistics (see section 2.1), we compute three summary statistics for each null hypothesis: a two-sided $p$-value, a BH $q$-value, and an lfdr estimate using the '`fdrtool`' package (Strimmer, 2008). These were 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 3 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 estimated lfdr is well-calibrated: among $t$-statistics for which the estimated lfdr is close to 25%, close to a quarter correspond to true null hypotheses. The calibration is even better for small values of estimated lfdr.

Local false discovery rates are also naturally derived from FDR quantities in the multiple

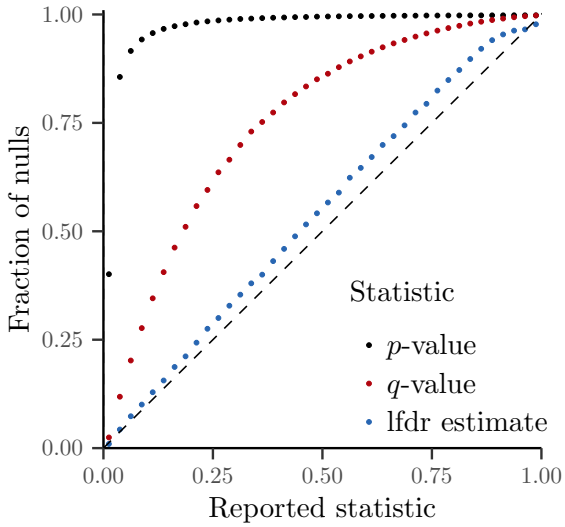


Figure 3: GGM example. Off-diagonal entries of $\Omega$ are either zero or $O\left(n^{-1/2}\right)$, where $d = 80$ and $n = 10d$. There are $m = \binom{d}{2} = 3,160$ tests, 95% of which are true nulls. We repeat the experiment $10^4$ times to assess calibration.

testing literature. The FDP and marginal FDR (mFDR) of a subset $A \subseteq \mathcal{Z}$ are defined:

$$\mathrm{FDP}(A) \coloneqq \frac{V(A)}{1 \vee R(A)}, \quad \mathrm{mFDR}(A) \coloneqq \frac{\mathbb{E}V(A)}{\mathbb{E}R(A)}$$

where

$$V(A) \coloneqq \#\{i : H_i \text{ is true}, z_i \in A\}$$
$$R(A) \coloneqq \#\{i : z_i \in A\}.$$

Similarly, the subset version of the positive FDR (pFDR) is defined

$$\mathrm{pFDR}(A) \coloneqq \mathbb{E}\left(\mathrm{FDP}(A) \mid z_i \in A \text{ for some } i\right).$$

Usually, FDR quantities are viewed as properties of a multiple testing procedure's entire rejection set. The above definitions generalize them to arbitrary subsets of the sample space. Our next result states that the frequentist lfdr function is equal to the limiting marginal or positive FDR within an interval shrinking to a point, provided that the joint density of the observations isn't supported on a zero-measure subset.

**Theorem 4.1.** *Suppose $z_1, \ldots, z_m$ are jointly absolutely continuous. Then, for any $t$*

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

The second equality in Theorem 4.1 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 (21) 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.

**Proposition 4.1.** *Suppose* $z_i \sim f^{(i)}$ *are independent, where* $f^{(i)} = f_0$ *when* $H_i = 0$ *and* $\varepsilon_m \to 0$ *is a sequence with* $m_0\varepsilon_m \to \infty$*, and that* $\text{lfdr} : \mathbb{R} \to [0,1]$ *is differentiable. Let* $\alpha \in \text{range}(\text{lfdr})$*. If* $|\text{lfdr}'(t)|$ *is bounded away from zero on* $\{t : \text{lfdr}(t) = \alpha\}$*, then*

$$\text{FDP}\big(\{z_i : |\text{lfdr}(z_i) - \alpha| \le \varepsilon_m\}\big) \xrightarrow{\mathbb{P}} \alpha$$

*as* $m \to \infty$*.*

Proposition 4.1 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 $(Y_1, \ldots, Y_{m-1})$. In our setting, the truth status of the null hypotheses are never revealed, and the sharpest way[2] to forecast whether $H_i$ is true depends on the observed $p$-values and the frequentist lfdr function. However, as Figure 3 demonstrates, estimates of the lfdr can also be approximately calibrated. In the next section, we turn our attention to estimating the lfdr.

## 5 Estimating the lfdr

Section 1 expressed the frequentist lfdr as the intensity ratio $\bar{\pi}_0 f_0(t)/\bar{f}(z)$. Assuming $f_0$ is known, we may conservatively bound $\bar{\pi}_0 \le 1$ or estimate it via e.g. Storey's method, reducing the problem of estimating lfdr 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 $z_1, \ldots, z_m \sim f$. Many parametric and nonparametric methods have been proposed to estimate $f$. Consider the maximum likelihood estimator

$$\hat{f}_m := \underset{f \in \mathcal{F}}{\text{argmax}} \ \frac{1}{m} \sum_{i=1}^{m} \log f(z_i), \tag{7}$$

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 (7) 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:

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

[2] among forecasters whose forecast about $H_i$ depends only on $z_i$

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 (7) matches the log-likelihood of the data. Nevertheless, the $M$-estimator $\hat{f}_m$ can still be computed from the sequence $z_1, \ldots, z_m$. 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 (7), 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 (7) is:

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

where $Z$ is a draw from the average density $\bar{f}$. Let $f_m^*$ denote the maximizer for the deterministic analog of (7)

$$\begin{aligned} f_m^* &\coloneqq \operatorname*{argmax}_{f \in \mathcal{F}} \mathbb{E}_{\bar{f}} \log f(Z) \\ &= \operatorname*{argmin}_{f \in \mathcal{F}} D(\bar{f} \parallel f), \end{aligned} \tag{8}$$

where $D(g\|h)$ 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 (7) converges uniformly to the population-level objective in (8), we will have $D(f_m^* \| \hat{f}_m) \xrightarrow{p} 0$ (Van der Vaart, 2000, Theorem 5.7). We record this observation in the following proposition.

**Proposition 5.1.** *Suppose* $M_m^*(f) \coloneqq \mathbb{E}_{\bar{f}} \log f(Z)$ *and* $\widehat{M}_m(f) \coloneqq \frac{1}{m}\sum_{i=1}^{m} \log f(z_i)$ *satisfy*

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

*as* $m \to \infty$ *and suppose that* $\bar{f} \in \mathcal{F}$. *Then* $D(\bar{f} \| \hat{f}_m) \xrightarrow{p} 0$.

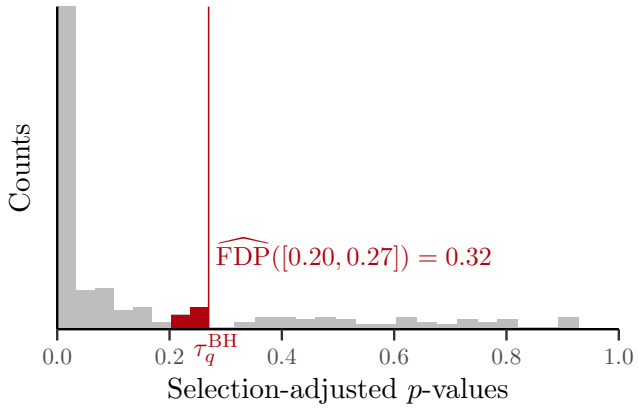


Figure 4: Shown above is the histogram of one-sided $p$-values falling below 0.025, adjusted for selection by multiplying by 40 (the reciprocal of the 2.5% one-sided significance threshold). The BH($q$) threshold for $q = 10\%$ is around 0.27 (or $\approx 0.007$ on the scale of the unadjusted $p$-values), below which there are 202 rejections. The estimated FDP near the edge of the rejection set (red) is around 32%.

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 $\mathcal{F} = \text{conv}(\mathcal{F}_0)$. 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.

In the next section, we illustrate methods for estimating the local false discovery rate on a meta-dataset collected by Mertens et al. (2022a) from the literature on 'nudges' in behavioral psychology.

### 5.1 Example: Analysis of nudge-data

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), Szaszi 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-

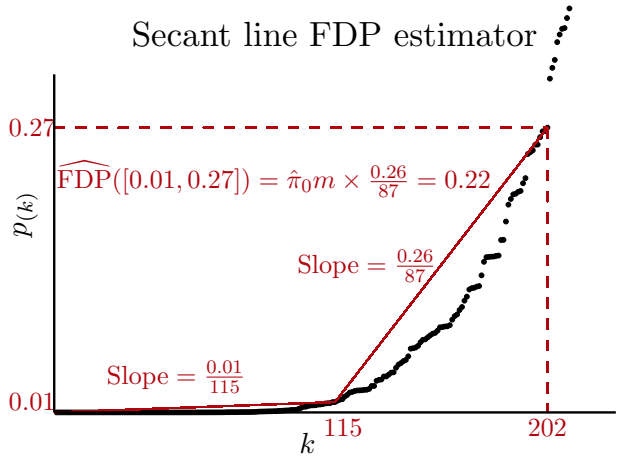


Figure 5: Nudge example. The order statistics of the selection-adjusted $p$-values are plotted against their rank. The estimated FDP among the smallest 202 $p$-values is close to 10%. Within the first half, the estimate is 0.6%, whereas in the second half it is 22%.

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)).

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 4, 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. As Figure 4 shows, the estimated proportion of false discoveries (FDP) grossly exceeds 10% for a subset of rejections near the rejection threshold. The method used to estimate the FDP is illustrated more explicitly in Figure 5.

The nudge example suggests that the BH rejection set is overly liberal in its last few rejections, which are of low quality compared to the rest. Instead of controlling the rate of false discoveries on average throughout the entire rejection set, we should focus on the FDR among $p$-values falling at or just below the rejection threshold. By controlling the quality of the least promising rejections, we ensure that each rejection is of sufficiently high quality. In the next section, we propose a new (frequentist) error criterion that puts this idea into practice.

## 6 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 $\mathcal{R} = \{i : \text{reject } H_i = 0\}$ according to its *least promising* rejection,

$$\text{max-lfdr}(\mathcal{R}) = \mathbb{E}\left[\max_{i \in \mathcal{R}} \mathbb{P}(H_i = 0 \mid p_i)\right].$$

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 the last rejection is a false discovery,

$$\text{bFDR}(\mathcal{R}) \coloneqq \mathbb{P}(H_{(R)} = 0), \tag{9}$$

where $H_{(0)} \coloneqq 1$ indicates the event where no rejections are made, and the notation $H_{(k)} \in \{0, 1\}$ means the hypothesis corresponding to the $k$th smallest $p$-value.

Definition (9) is at first glance puzzling; if the hypotheses $H_1, \ldots, H_m \in \{0, 1\}$ are fixed to begin with, then how can we speak of the probability that one of them is null? The reason is that we are not asking about a fixed and preconceived hypothesis, but a random one depending on the data.

### 6.1 Comparison with FDR

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

$$\text{FDR}(\mathcal{R}) = \mathbb{P}(H_{(I)} = 0), \quad I \sim \text{Uniform}\{1, \ldots, R\}.$$

Figure 6 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 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.

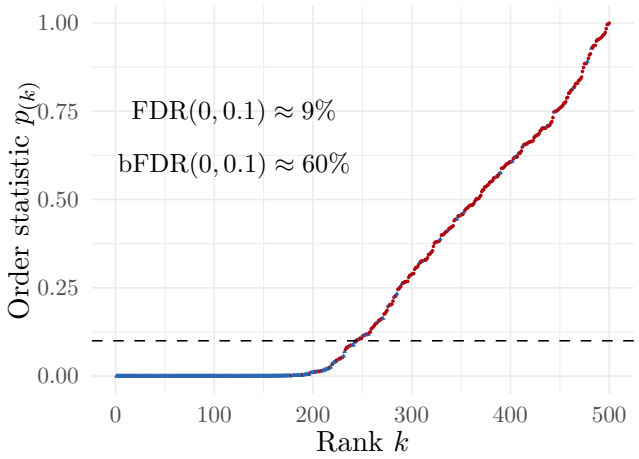


Figure 6: The order statistics of $m = 500$ $p$-values are plotted against their rank. They are generated with $\bar{\pi}_0 = 0.5$ where null $p$-values (red) are i.i.d. Uniform$(0, 1)$ and alternative $p$-values (blue) are i.i.d. Beta$(0.05, 1)$. The bFDR is approximated by the fraction of nulls among the largest 15 $p$-values below 0.1.

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

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

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.

## 6.2 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 $\alpha$ rejects the $R_\alpha$ smallest $p$-values, where

$$R_\alpha \coloneqq \underset{k=0,\ldots,m}{\operatorname{argmax}} \left\{ \frac{\alpha k}{m} - p_{(k)} \right\}, \quad p_{(0)} \coloneqq 0. \tag{10}$$

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

**Theorem 6.1.** *If $p_1, \ldots, p_m$ are independent and $H_i = 0$ implies $p_i \sim \text{Uniform}(0,1)$, then*

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

*where $\mathcal{R}_\alpha \coloneqq \{i : p_i \leq p_{(R_\alpha)}\}$ is defined by (10).*

*Proof of Theorem 6.1.* The event $\{H_{(R_\alpha)} = 0\}$ can be written as a disjoint union,

$$\{H_{(R_\alpha)} = 0\} = \bigcup_{i:H_i=0} \{p_{(R_\alpha)} = p_i\}$$

which implies

$$\mathbb{P}(H_{(R_\alpha)} = 0) = \sum_{i:H_i=0} \mathbb{P}(p_{(R_\alpha)} = p_i) = m_0 \cdot \frac{\alpha}{m}.$$

The last equality follows from Lemma 2 of Soloff et al. (2024), which states that for any configuration of the other $p$-values $p_1, \ldots, p_{m-1}$, the probability that a null $p$-value $p_m$ achieves the optimum in (10) is equal to $\frac{\alpha}{m}$[3]. □

The proof of Theorem 6.1 shows the boundary FDR control of the SL procedure is controlled when each null density is bounded,

$$H_i = 0 \Rightarrow f^{(i)}(t) \leq 1 \quad \text{for all } t \in [0, \alpha]. \tag{11}$$

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 control. An example in which super-uniformly distributed $p$-values arise is one-sided Gaussian location testing,

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

where $H_i = 0 \Rightarrow \theta_i \leq 0$. In this case, the probability density function for $p_i = 1 - \Phi(X_i)$ satisfies (11) under the null, for any $\alpha \leq 1/2$. This observation extends to one-parameter exponential families with continuous densities.

**Proposition 6.1.** *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 $(G_\theta)$. 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 $\theta_0$. 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 $\theta \leq \theta_0$.*

[3] An alternative proof of the fact "$H_i = 0 \Rightarrow \mathbb{P}(p_{(R_\alpha)} = p_i) = \frac{\alpha}{m}$" can be found in the appendix, the technical key for which is a telescoping sum argument.

# 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. Frequentists 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 5.1), 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? We thank Bradley Efron, Rina Foygel Barber, Chao Gao, Ruth Heller, Peter McCullagh, Asaf Weinstein and Dani Yekutieli 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. William Fithian was supported by the NSF DMS-1916220 and a Hellman Fellowship from Berkeley.


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

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

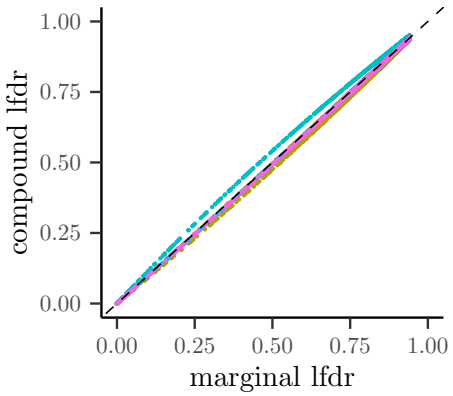


Figure 7: For each of 6 realizations of the vector $(p_1, \ldots, p_m)$, with $m = 1000$, $\bar{\pi}_0 = 0.8$, $f_0 = 1_{[0,1]}$ and $f_1 = \text{Beta}(1/4, 1)$, $\text{clfdr}(\mathbf{p})$ is approximated numerically and the points $(\text{lfdr}(p_i), \text{clfdr}_i(\mathbf{p}))$ are plotted with the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

# 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 $p_1, \ldots, p_m$. 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 $\delta(\mathbf{p}) \coloneqq (\delta_1(\mathbf{p}), \ldots, \delta_m(\mathbf{p}))$ is is permutation equivariant (PE) if

$$\delta(\mathbf{p})_\pi = \delta(\mathbf{p}_\pi) \quad \text{for any } \pi \in \mathcal{S}_m, \tag{12}$$

where $\mathcal{S}_m$ is the set of permutations on $[m]$, and $v_\pi \coloneqq (v_{\pi(1)}, \ldots, v_{\pi(m)})$ denotes the vector $v \in \mathbb{R}^m$ permuted by $\pi$. 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:

$$\begin{aligned} \pi &\sim \text{Uniform}(\mathcal{S}_m) \\ \widetilde{H} &\coloneqq H_\pi, \quad \widetilde{\mathbf{p}} \coloneqq \mathbf{p}_\pi. \end{aligned} \tag{13}$$

The weighted classification risk of any PE decision rule $\delta$ in this model coincides with its frequentist compound risk, yielding another instance of the fundamental theorem for compound decisions (Zhang (2003), Robbins (1951), Weinstein (2021))

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

where $\widetilde{\mathbb{E}}$ marginalizes over $\widetilde{H}$ and $\widetilde{\mathbf{p}}$ generated by (13). The right hand side is minimized by the Bayes rule within the exchangeable oracle model (13), characterized by the compound lfdr (clfdr),

$$\begin{aligned} \text{clfdr}_i(t_1, \ldots, t_m) &\coloneqq \mathbb{P}(\widetilde{H}_i = 0 \mid \widetilde{\mathbf{p}} = \mathbf{t}) \\ &= \frac{\sum_{\pi \in \mathcal{S}_m : H_{\pi(i)} = 0} \prod_{j=1}^m f^{(\pi(j))}(t_j)}{\sum_{\pi \in \mathcal{S}_m} \prod_{j=1}^m f^{(\pi(j))}(t_j)}, \end{aligned} \tag{14}$$

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

$$\delta_i^*(\mathbf{p}) \coloneqq \begin{cases} 1 & \text{if clfdr}_i(\mathbf{p}) \leq \frac{1}{1+\lambda} \\ 0 & \text{else.} \end{cases} \tag{15}$$

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 (13) by conditioning on one $p$-value,

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

Given the true $p$-value densities $f^{(1)}, \ldots, f^{(m)}$, 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 lfdr is a fixed function on $[0,1]$, clfdr depends on the particular realization of $p$-values, as illustrated in Figure 8. The clfdr and lfdr 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 lfdr 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 lfdr 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,

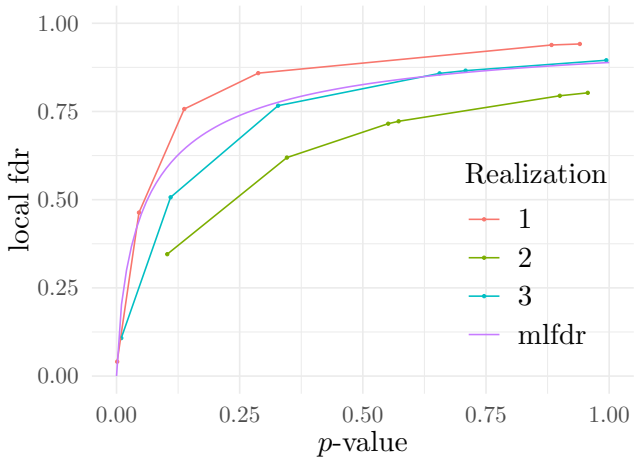


Figure 8: Three realizations of the order statistics of $m = 6$ many $p$-values are plotted against their scores $\text{clfdr}_i(\mathbf{p})$ for $i = 1, \ldots, 6$. In this simulation, $m_0 = 4$, $f_0 = \text{Uniform}(0, 1)$, and $f_1 = \text{Beta}(1/4, 1)$. The clfdr scores are computed using the realized values $p_{(1)} \leq \cdots \leq p_{(6)}$ and formula (14). The dependence between clfdr scores in any given realization requires that they always sum to $m_0$.

suppose there is an exchangeable sequence of latent variables $\theta_1, \ldots, \theta_m$ taking values in some parameter space $\Theta$, and conditional on $\theta = (\theta_1, \ldots, \theta_m)$, the data is drawn according to

$$p_i \mid \theta \sim f_{\theta_i}, \quad \text{independently for } i = 1, \ldots, m. \tag{16}$$

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 $\phi$ 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 $\theta$, the marginal and compound lfdr in a Bayesian model with an exchangeable prior on $\theta$ are:

$$\text{lfdr}(t; G_m) = \mathbb{P}(\theta_i = 0 \mid p_i = t, G_m), \tag{17}$$

$$\text{clfdr}_i(\mathbf{t}; G_m) = \mathbb{P}(\theta_i = 0 \mid \mathbf{p} = \mathbf{t}, G_m) \tag{18}$$

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

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

is the empirical cumulative distribution function of the true effects, and $\{\theta_i = 0\}$ is the null event[5]. 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 $\theta$, the joint distribution (16) defines a frequentist model, and within

[5] previously denoted $H_i = 0$

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 $\theta_1, \ldots, \theta_m$ is an exchangeable sequence of latent variables, and that given $\theta = (\theta_1, \ldots, \theta_m)$, the p-values are drawn according to (16). Let $\mathrm{lfdr}(t; G_m)$ and $\mathrm{clfdr}_i(\boldsymbol{t}; G_m)$ be defined as in (17) and (18) where $\boldsymbol{t} \coloneqq (t_1, \ldots, t_m) \in [0,1]^m$. Then*

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

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

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 (16) holds for each $i = 1, \ldots, m$. Then*

$$\mathbb{P}(\theta_i = 0 \mid \boldsymbol{p}) = \mathbb{E}\left[\mathrm{clfdr}_i(\boldsymbol{p}; \theta) \mid \boldsymbol{p}\right]. \tag{19}$$

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

$$\mathbb{P}(\theta_i = 0 \mid p_i) = \mathbb{E}\left[\mathrm{lfdr}(p_i; \theta) \mid p_i\right]. \tag{20}$$

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 $m \to \infty$.

**Lemma A.1.** *Suppose $p_i \sim f^{(i)}$ are drawn independently for $i = 1, \ldots, m$ 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 $\frac{m_0}{m} \to \pi_0 \in (0,1)$*

*as* $m \to \infty$, *and* $\mathrm{Var}\left(\frac{f_1}{f_0}(p_1)\right) \vee \mathrm{Var}\left(\frac{f_0}{f_1}(p_2)\right) < \infty$ *when* $p_1 \sim f_0$ *and* $p_2 \sim f_1$, *then we have for each* $i = 1, 2, \ldots$

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

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

# B Proofs of technical results

*Proof of Theorem 3.1.* 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,

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

*Proof of Theorem 3.2.* The continuous density assumption implies

$$\mathbb{P}\Big(\bigcup_{j \in \mathcal{H}_0} \{|\ell_j - \alpha| \le \varepsilon\}\Big) \sim \sum_{j \in \mathcal{H}_0} \int_{\{t : |\mathrm{lfdr}(t) - \alpha| \le \varepsilon\}} f_0(t) dt$$
$$\mathbb{P}\Big(\bigcup_{j=1}^m \{|\ell_j - \alpha| \le \varepsilon\}\Big) \sim m \int_{\{t : |\mathrm{lfdr}(t) - \alpha| \le \varepsilon\}} \bar{f}(t) dt,$$

as $\varepsilon \to 0$. Now since $\bar{\pi}_0 f_0(t) \sim \alpha \bar{f}(t)$ for any $t$ such that $|\mathrm{lfdr}(t) - \alpha| \le \varepsilon$, the ratio tends to $\alpha$.

Since $g$ is calibrated,

$$\begin{aligned} g(t) &= \mathbb{E}\left[1 - H_I \mid g(z_I) = g(t)\right] \\ &= \mathbb{E}\left[\mathrm{lfdr}(z_I) \mid g(z_I) = g(t)\right], \end{aligned}$$

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

*Proof of Theorem 3.3.* 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 $I \sim \mathrm{Uniform}\{1, \ldots, m\}$ and $\ell_\lambda(h, y)$ is the per-instance loss:

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

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,

$$\mathrm{lfdr}(t) = \mathbb{P}(H_I = 0 \mid z_I = t). \tag{21}$$

$\square$

*Proof of Theorem 4.1.* Let $V \equiv V(z-\varepsilon, z+\varepsilon)$ and $R \equiv R(z-\varepsilon, z+\ \varepsilon)$. First, mFDR($[t-\varepsilon, t+\varepsilon]$) is equal to

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

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

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

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

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

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

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

□

*Proof of Proposition 4.1.* Let $S_m \coloneqq \{t \in [0,1] : |\mathrm{lfdr}(t) - \alpha| \le \varepsilon_m\}$, and define the count variables

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

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

$$\frac{N_0}{\mathbb{E}N_0} - 1 \xrightarrow{\mathbb{P}} 0, \;\text{ and }\; \frac{N_0+N_1}{\mathbb{E}(N_0+N_1)} - 1 \xrightarrow{\mathbb{P}} 0, \tag{22}$$

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

$$\mathbb{P}\left(\left|\frac{N_0}{\mathbb{E}N_0} - 1\right| > \delta\right) \le \frac{\mathrm{Var}(N_0)}{(\mathbb{E}N_0)^2\delta^2} \asymp \frac{m_0\varepsilon_m}{(m_0\varepsilon_m)^2}, \tag{23}$$

because $H_i = 0$ implies $\mathbb{P}(\text{lfdr}(z_i) \in \alpha \pm \varepsilon_m) \asymp \varepsilon_m$ as $m \to \infty$, since for each $\alpha$ 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 (23) tends to zero since $m_0\varepsilon_m \to \infty$. Similarly,

$$\mathbb{P}\left(\left|\frac{N_0+N_1}{\mathbb{E}(N_0+N_1)} - 1\right| > \delta\right) \le \frac{\text{Var}(N_0+N_1)}{(\mathbb{E}(N_0+N_1))^2\delta^2} \asymp \frac{m\varepsilon_m}{(m\varepsilon_m)^2} \le \frac{1}{m_0\varepsilon_m} \to 0,$$

from which (22) follows. □

*Proof of Proposition 5.1.* By Theorem 5.7 in Van der Vaart (2000), it suffices to check that $f_m^*$ is well-separated, i.e. for every $\varepsilon > 0$,

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

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

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

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

*Alternative proof of Theorem 6.1.* Suppose without loss of generality that $H_m = 0$. Then since the nulls are exchangeable, the bFDR of the SL method is

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

Let $q_{(1)} \le \cdots \le q_{(m-1)}$ denote the order statistics of $p_1, \ldots, p_{m-1}$, and note that $p_m$ achieves the maximum in (10) as the $(k+1)$th order statistic if $q_{(k)} < p_m < q_{(k+1)}$ and

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

for $k \le m-1$, where $q_{(0)} \coloneqq 0$ and $\Delta_j \coloneqq \frac{\alpha j}{m} - q_{(j)}$. 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

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

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

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

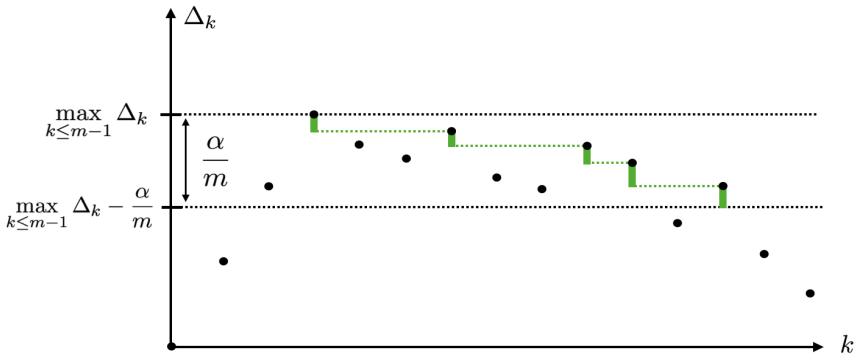


Figure 9: Each length of the interval range in which $p_m$ achieves the maximum in (10) is indicated by a vertical green bar, and the sum of these lengths is $\frac{\alpha}{m}$.

The sum of lengths of the non-empty ranges is telescoping and equal to $\frac{\alpha}{m}$, as illustrated in Figure 9. □

*Proof of Proposition 6.1.* When $Z \sim g_\theta$, the density of $p = 1 - G_{\theta_0}(Z)$ is

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

At $\theta = \theta_0$, the above ratio is equal to 1. When $\theta \le \theta_0$, the log density has a positive derivative in $\theta$ 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 $t \le \alpha$ if $G_{\theta_0}^{-1}(1-\alpha) > \mathbb{E}_{\theta_0}(Z)$. □

**Proposition B.1.** *In the setting of section A.1, the best* PE *decision rule for minimizing the weighted classification risk is defined by* (14) *and* (15).

*Proof.* For any PE decision rule $\delta$,

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

The Bayes rule

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

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

$$(\widetilde{H}, \widetilde{\mathbf{p}}) \overset{(d)}{=} (\widetilde{H}_\sigma, \widetilde{\mathbf{p}}_\sigma),$$

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

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

Since the above holds for any permutation $\sigma$, the Bayes rule in model (13) is a PE decision rule. Since the average risk in the Bayes model (13) 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

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

*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 $\theta_{(1)} \leq \cdots \leq \theta_{(m)}$ are the ordered values of $\theta_1, \ldots, \theta_m$. Since $\theta_1, \ldots, \theta_m$ 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 (18), note that when $p_i = p_{(k)}$, exchangeability implies

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

□

*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 $H_2 = 1$,

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

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:

$$\begin{aligned}
&\mathrm{Hyp}_0 : \text{Observe a random permutation of } p_{-i} \text{ when } H_i = 1 \\
&\mathrm{Hyp}_1 : \text{Observe a random permutation of } p_{-i} \text{ when } H_i = 0,
\end{aligned}$$

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

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

since $\mathrm{Hyp}_0, \mathrm{Hyp}_1$ can be obtained from $\widetilde{\mathrm{Hyp}}_0, \widetilde{\mathrm{Hyp}}_0$ 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 $\mathrm{Hyp}_1$ (resp. $\mathrm{Hyp}_0$). The likelihood ratio of $\widetilde{\mathrm{Hyp}}_1$ to $\widetilde{\mathrm{Hyp}}_0$ has variance

$$\mathrm{Var}_0 \left( \frac{1}{m_0} \sum_{\ell=1}^{m_0} \frac{f_1}{f_0}(\widetilde{p}_\ell) \right) = \frac{1}{m_0} \mathrm{Var}_0 \left( \frac{f_1}{f_0}(p_1) \right) \to 0$$

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

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

A symmetric argument yields

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

under the condition that $\mathrm{Var}\left(\frac{f_0}{f_1}(p_2)\right) < \infty$ when $H_2 = 1$. Here we are abusing notation by writing the index $\ell$ 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 $\widetilde{\mathrm{Hyp}}_1$. It now follows from Chebyshev's inequality that

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

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

## 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*

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*

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., 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* .

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.