False Discovery Rates: Algorithms and Estimations (In Progress)

Wen-Hung Wang
Advisor: Chun-Hao Yang

November 29, 2022

Abstract

This study surveys existing research results concerning false discovery rates (FDRs), aiming to give readers a picture of the development of multiple comparison procedures (MCPs) controlling FDRs and their properties as well as how these MCPs can be used. We first discuss the motivation for FDRs and then present MCPs including the Benjamini and Hochberg's procedure, an adaptive method from Liu and Sarkar (2011), MCPs using the q-value, and an MCP using the local FDR. The emphasis of the study is on the estimations of FDRs under independent test statistics. A brief discussion on dependent test statistics is given as well. Applications are discussed with a concentration on a labor study. Finally, we empirically explore the properties these MCPs by numerical experiments under various settings [This part is still in progress].

1 Introduction

A hypothesis is a statement of a population parameter that consists of a null hypothesis H_0 and an alternative hypothesis H_1, which is complementary of H_0. A testing procedure rejects the null based on some rule and in the same time controls the Type I error rate at level α, the probability of falsely reject the null, i.e., Pr(reject $H_0|H_0$ is true). For example, we would like to test if μ is 0 in $N(\mu, \sigma^2)$ with unknown σ^2. We can simply test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$ based on a t statistic T. One rejects $H_0 : \mu = 0$ if $|T| > t$ for some t such that the Type I error is α.

In the situations similar to the above test containing only a statement H_0 versus H_1, the Type I error can be defined simply as Pr(reject $H_0|H_0$ is true). However, in some applications, one wishes to test many statements "simultaneously," i.e., test a set of hypotheses $\{H_{0i}$ versus $H_{1i}\}_{i=1}^m$ simultaneously, which is called a multiple test or multiple comparison, where H_{i0} and H_{i1} are the null and the alternative of hypothesis i, respectively. These applications include testing a group of gene expression level differences between normal and cancered cells, testing multiple endpoints in a clinical trial in new drug development, and among others. The set $\{H_{0i}$ versus $H_{1i}\}_{i=1}^m$ is often called a "family" of tests.

A conventional evaluation metric in multiple tests is called the familywise error rate (FWER), which is defined to be the probability of incurring one or more false rejection(s), i.e., $\Pr(V \geq 1)$ from Table 1. A rejection is sometimes called a discovery. A multiple comparison procedure (MCP) is the testing procedure that controls the Type I error of a multiple test at some $\alpha \in [0, 1]$ level. An MCP assigns a threshold to an individual test H_i so that the Type I error rate of the whole multiple test can be controlled. If one naively thresholds each test i at α, then the resulting FWER is $1-(1-\alpha)^m$, which is greater than α. In such a situation, the FWER is not

	Not Significant	Significant (Discoveries)	Total
True Null	U	V	m_0
Non-Null	T	S	m_1
	$m - R$	R	m

Table 1: Relevant random variables defined through out the article. m is known and R is observable. Others are unobservable.

controlled. This problem is called a "multiplicity" problem in multiple testing. Hence some type of correction for individual thresholds is required, e.g., a Bonferroni's correction α/m. A threshold stricter than α is needed to adjust for multiplicity. Without the adjustment for multiplicity in a multiple test, the resulting inference could be misleading, e.g., making erroneous conclusions about a drug's effects [Food and Drug Administration (2022)].

Although the FWER can be used in multiple testing, Benjamini and Hochberg (1995) argued that this error rate is so stringent that it sometimes yields a low power, the probability of finding true discoveries in some situations, especially when m is large. They further contended that guarding against one false discovery like the FWER is not necessary in some applications whose goals are to find as many discoveries as possible. A motivating example from Example 1 in Storey and Tibishirani (2003) is finding differently expressed genes in a microarray experiment from thousands of genes. In this case, m could be quite large, leading to only a small number of rejections if we use an MCP controlling the FWER.

Inspired by those applications, Benjamini and Hochberg proposed a new notion of error rate in 1995: an error rate that is scalable so that it can be properly applied when m is large. They defined the false discovery rates (FDR) as the expected proportion of false discoveries to all rejections. Precise definition of FDR is given in section 2.1. A desirable property of FDRs is that it is scalable: When the number of rejections becomes large, the rate allows more false discoveries as well. This gives an improvement of power in large scale tests where m is large since it is less strict. As a result, MCPs controlling FDRs is favored over those controlling the FWER in a high-dimensional test.

This article aims to give the reader a picture of the development of MCPs controlling FDRs and their properties as well as how these procedures can be exploited in practice. Such MCPs can be derived from two perspective: algorithm or estimation. We mainly focus on the estimation side of FDRs. Two algorithm-derived MCPs are given in section 2, including the standard Benjamini and Hochberg's procedure and an adaptive method from Liu and Sarkar (2011). In section 3, MCPs developed from estimation viewpoints are introduced. Two estimators and their properties are addressed. The concept of the q-value is introduced as well.

Another FDR called the local FDR is discussed in section 4. We discuss the motivation for defining this error rate, a special transformation and basic setupsm, and estimation methods from Efron (2007) and Efron (2012). Power, influence analyses, as well as the comparison of the q-value and the local FDR are included.

In section 5, we introduce some works regarding dependence of test statistics [Yekutieli and Benjamini (1999), Schwartzman and Lin (2011), hang, Fan and Yu (2011), etc]. An application of the q-value procedure in a labor study from Kilne, Rose and Walters (2022) is discussed in section 6. Some numerical experiments will be reported in section 7 [This part is still in progress].

2 Algorithms

In the development of MCPs with the error rates being FDRs, there are two approaches from an algorithm perspective or an estimation perspective. The former is discussed in this section

and the latter will be introduced in section 3. The algorithm perspective fixes an acceptable significance α "beforehand" and finds thresholds applied to each individual tests, e.g., $(i/m)\alpha$ for test i in Benjamini and Hochberg's Procedure. Essentially, an algorithm estimates the rejection region of an MCP. Section 2.1 and 2.2 provide two MCPs controlling FDR, which are developed in an algorithm view .

2.1 Benjamini and Hochberg's Procedure

The Benjamini and Hochberg's procedure (the BH procedure) is a standard MCP that controls the FDR at a predetermined level α and was derived from an algorithm perspective. In Benjamini and Hochberg (1995), they first considered different formulations of FDR:

$$E\left[\frac{V}{R \vee 1}\right] = E\left[\frac{V}{R}\middle| R > 0\right] \Pr(R > 0) \tag{1}$$

$$E\left[\frac{V}{R}\middle| R > 0\right] \tag{2}$$

$$\frac{E[V]}{E[R]} \tag{3}$$

The values of (2) and (3) will be 1 if $m = m_0$, i.e., all discoveries are false, the error rates can not be controlled under α when $\alpha < 1$, when is usually the case. Therefore, under an algorithm's viewpoint, (1), which is less than 1, is more attractive. Interestingly, in the context of estimating FDRs, (2) has some well-behaved properties and interpretations, see Storey (2003). In literature, (2) is called the positive FDR (pFDR) that will be discussed later. For tests in which m is large, these three quantities are equivalent.

As stated previously, the FDR is less stringent than the FWER; thus MCPs controlling the FDR are expected to have greater power than those controlling the FWER. It can be shown that for $m = m_0$, the FDR is equal to the FWER; for $m_0 < m$, the FDR is equal to or less than the FWER. Hence the larger $m_1 = m - m_0$ is, the more gain in power using MCPs controlling the FDR over those controlling the FWER one will have.

Benjamini and Hochberg adopted the sequential p-value procedure from Sime's approach that controls the FWER. The resulting method is the BH procedure, which is a standard MCP controlling the FDR [Section 3.1, Benjamini and Hochberg (1995)]. Let $H_1, ..., H_m$ be the hypotheses that we wish to test simultaneously, with corresponding independent p-values $P_1, ..., P_m$. Let $P_{(1)} \leq ... \leq P_{(m)}$ be the ordered p-values that corresponds to tests $H_{(1)}, ...H_{(m)}$. The BH procedure works as follows: reject all $H_{(i)}, i = 1, ..., k$, where $k = \max\{i : P_{(i)} \leq (i/m)\alpha\}$ if k exists; otherwise rejects nothing. They proved that this MCP satisfies:

$$\text{the FDR of the BH method} \leq \frac{m_0}{m}\alpha \leq \alpha. \tag{4}$$

Some remarks can be made so far:

1. The procedure requires p-values to be independent. In fact, in their proof, the independence assumption is only needed for null p-values, P_i's from true null H_i's. Many subsequent works develop MCPs under dependence to relax the independence restriction, as will be addressed in section 5.

2. The proof in Benjamini and Hochberg (1995) relies on the fact $\Pr(P_i \leq u \mid H_i \text{ is truly null}) = u$, i.e., p-values will be uniformly distributed between (0, 1) under null. This holds only for continuous P_i's [$\Pr(P_i \leq u \mid H_i \text{ is truly null}) \leq u$ for discrete P_i's]. Benjamini and Yekutieli (2001) provided a more general proof of the BH procedure for both continuous and discrete P_i's [Section 5, Benjamini and Yekutieli (2001)].

3. For continuous P_i's the BH procedure controls the FDR level at "exact" $(m_0/m)\alpha$, whereas for discrete P_i's it controls the rate less or equal to $(m_0/m)\alpha$ [Theorem 5.1, Benjamini and Yekutieli (2001)].

4. The inequality in (4) can be further sharpened, which could lead to a potential improvement in power by including more false discoveries within the same tolerance level. To see this, suppose we know m_0, the number of truly null hypotheses, and apply $(i/m_0)\alpha$ to threshold the ordered p-values, then we can control the FDR level at $(m_0/m_0)\alpha = \alpha$ in (4), i.e., the gap between $(m_0/m)\alpha$ and α disappears. However, we will never know m_0. Alternatively, many researchers attempted to incorporate the information of m_0 into their algorithms or estimations to increase the power of their MCPs controlling FDRs, as will be present in section 2.2, 3.1, and 4.3 [Liu and Sarkar (2011), Storey (2002), Efron (2007), etc].

5. The BH procedure does not need the information of the distribution of non-null p-values. This is a unique property due to the choice of its threshold $(i/m)\alpha$ [Theorem 5.3, Benjamini and Yekutieli (2001)].

The numerical studies in Benjamini and Hochberg (1995) show that in terms of power, the BH procedure is powerful than the Bonferroni's and Hochberg's (1988) procedures. The advantages increase as m_1 or m becomes large, which is expected. When m becomes larger, the power declines. This phenomena, they claimed, is the price of multiplicity control. However, the MCP derived from the estimator of Storey (2002) shows the opposite: the power of the MCP increases as m becomes large.

2.2 An Adaptive Control Algorithm

From the observation that replacing m with m_0 in the BH procedure sharpens inequality (4), it is tempting for one to take advantage of this property to improve the MCP. Here, we present a multi-stage algorithm that first estimates m_0 to incorporate the information of m_0. The MCP proposed by Liu and Sarkar in 2011 uses an algorithm perspective to obtain $\hat{m}_0$ as an estimate of m_0.

Let α_i be the individual threshold for test i. A step-up procedure rejects $H_{(i)}$, $i = 1, ..., k$ with $k = \max\{i : P_{(i)} \leq \alpha_i\}$ if k exists; otherwise rejects nothing. In contrast, a step-down procedure rejects $H_{(i)}$, $i = 1, ..., k$ with $k = \max\{i : P_{(j)} \leq \alpha_j, \text{ for all } j \leq i\}$ if k exists; otherwise rejects nothing. A single-step procedure thresholds each P_i at a constant $c \in [0, 1]$.

The proposed algorithm in Liu and Sarkar (2011) combines both step-down and step-up procedures in two stages: the first stage uses a step-down method to obtain $\hat{m}_0$; then the second stage replace m with $\hat{m}_0$ in the step-up BH method. They also proved that under p-values are independent the MCP controls the FDR at α [Procedure 1.1 and Theorem 1.1, respectively, Liu and Sarkar (2011)]. Their simulation studies indicate that under moderately positive correlation between p-values, the MCP still controls the FDR at α.

3 Estimations

This section introduces an estimation perspective, i.e., convert the MCP problem into estimations of FDRs. The estimations of FDRs will encounter some difficulties if one treats the MCP as an "algorithm," e.g., how to estimate the error rate under different thresholds? Therefore, to avoid such difficulties, one can often start with "a fixed significance region Γ" for all tests, an opposite approach to an algorithm type procedure. [Recall that an algorithm estimates those thresholds in tests].

Storey (2003) proposed deriving an MCP as follows. First fix the significance threshold t for all tests and then conservatively estimate the error rate, e.g., $\widehat{\text{FDR}} \geq \text{FDR}$, $E[\widehat{\text{FDR}}] \geq \text{FDR}$, etc [denote the BH's FDR as FDR for the rest of section 3 for simplicity]. Then utilize the obtained estimator to compute the q-values to conduct an MCP. This allows one to decide the acceptable significance level "after" estimation, as will be discussed in section 3.2. The main estimators we will cover are given by Storey (2002).

Under an algorithm's viewpoint one is forced to use the error rate in (1), the FDR in BH procedure, since (2) and (3) cannot be controlled if $m = m_0$. However, in the context of estimation, the positive FDR (2) possesses good properties. Therefore, we start with introducing this quantity. Let us note that the following discussion also involves the FDR (1) because asymptotically pFDR $\approx$ FDR [In the applications of FDR, large m is usually the case, i.e., high-dimensional tests].

Section 3.1 focuses on the FDR and positive FDR, which are "global" measurements of errors. Nonetheless, one might be curious about the significance of an individual test, i.e., a "local" measurement. The q-value in section 3.2 tries to give such information [so does the local FDR in section 4]. Section 3.2 first discusses an MCP using the q-value and relevant properties. The whole section 3 ends in explaining the relationship of estimation and algorithm. The methods in this section is provided in R package [Storey, Andrew, Dabney and Robinson (2015)].

3.1 Positive FDR (pFDR)

The positive FDR (pFDR) is defined in (2) as $\text{pFDR} = E[V/R|R > 0] = \text{the FDR}/\Pr(R > 0)$. In some extreme case whose $\Pr(R > 0)$ is small, say 0.5, these two quantities are quite different. Such a situation is more likely to happen in a "finite m" test [m is not large enough] than in a "large m" test where $\Pr(R > 0) \approx 1$. This gives the ground to make finite adjustments in estimations [Storey (2002)].

3.1.1 Properties of the pFDR

Storey (2003) gives a thorough theoretical investigation in pFDR. Here, we present some of the results and will show how subsequent works apply these properties into estimation in section 3.1.2 [Storey (2002), Storey and Tibshirani (2003), etc].

Let H_i, $i = 1, ..., m$ be m identical hypotheses to be tested, Γ be the fixed rejection region for all H_i, and $T_i, ...T_m$ be the test statistics. Without loss of generality, H_i is 0 if it is null; 1 if non-null, i.e., simple versus simple hypotheses [for composite non-null, the non-null distribution is a mixture]. Assume that test statistics come from a Bayesian two-class model

$$T_i|H_i \overset{i.i.d.}{\sim} (1 - H_i)F_0 + H_iF_1, \tag{5}$$

$$H_i \sim Bernoulli(\pi_1), \tag{6}$$

where F_0 is some null distribution, F_1 is some non-null distribution and $\pi_1 := \Pr(H_i = 1) := 1 - \pi_0$ for $i = 1, ..., m$. That is, (5) is the conditional likelihood of T_i and (6) is the prior probability of H_i. Under this setup, Storey (2003) Theorem 1 shows that

$$\text{pFDR}(\Gamma) := E\left[\frac{V(\Gamma)}{R(\Gamma)}\middle| R(\Gamma) > 0\right] = \Pr(H = 0|T \in \Gamma), \tag{7}$$

where $V(\Gamma) = \#\{T_i : T_i|H_i = 0, T_i \in \Gamma\}$, $R(\Gamma) = \#\{T_i : T_i \in \Gamma\}$, i.e., the pFDR obtained from rejecting test i if $T_i \in \Gamma$ is the posterior of the Bayesian two-class model in (5) and (6) [H_i, T_i can be replaced by H, T due to the i.i.d. assumption]. This result holds only under the

i.i.d. assumption on $T_i|H_i$. When dependence between T_i's emerges, the result no longer holds [Theorem 3, Storey (2003)]. However, when the empirical null probability, empirical p-value, and empirical power all converge, FDR and pFDR converge to some asymptotic posterior. This aymptotic condition is called weak dependence. [Theorem 4, Storey (2003)].

3.1.2 Estimation of pFDR and the FDR

With the property (7) and the relationship of $\{T_i \in \Gamma\} = \{P_i \leq t\}$ for some threshold $t \in [0, 1]$ where P_i can be obtained by permutation methods, Storey (2002) proposed estimators for the pFDR and the FDR for "independent" p-values P_i [(7) requires independence of test statistics] by the "zero assumption," which presumes that most P_i's near 1 will be truly null cases. This assumption is more realistic if π_0 is large, i.e., a sparse multiple test, and allows one to incorporate the information of m_0 by a "point" estimate of π_0 [Recall that the adaptive method in section 2.2 obtains $\hat{m}_0$ by an algorithm procedure]. Fig 1 of Storey and Tibshirani (2003) illustrates the assumption: observed p-values distribute uniformly when they are large, i.e., the flat area in the right-hand side of the histogram of observed p-values. A tuning parameter $\lambda \in [0, 1]$ reflects the assumption: p-values larger than λ are believed to be from null; and hence are distributed uniformly $\Pr(P_i \leq t|H_i \text{ is truly null}) = t$ [this result only holds for continuous p-values, for the discrete cases, $\Pr(P_i \leq t|H_i \text{ is truly null}) \leq t$, the p-values needs to be adjusted so that the below estimators can be used properly, Storey and Tibshirani (2003)].

As a result, Storey (2002) estimates $\text{pFDR}(t) = \text{pFDR}(\{P_i \leq t\})$ and $\text{FDR}(t) = \text{FDR}(\{P_i \leq t\})$ by the relationship

$$\text{pFDR}_\lambda(t) = \frac{\pi_0 \Pr(P \leq t|H = 0)}{\Pr(P \leq t)} = \frac{\pi_0 t}{\Pr(P \leq t)}, \tag{8}$$

where the last equality comes from $\Pr(P \leq t|H = 0) = t$. A naive estimate yields $\widehat{\Pr}(P \leq t) = \#\{p_i \leq t\}/m := R(t)/m$ with p_i are observed p-values and $R(t) = \#\{p_i \leq t\}$ for all i. Under the zero assumption, π_0 can be estimated: $\#\{p_i > \lambda\}$ is approximately $m\pi_0(\lambda)(1 - \lambda)$. Thus $m\hat{\pi}_0(\lambda)(1 - \lambda) = \#\{p_i > \lambda\}$, i.e., $\hat{\pi}_0(\lambda) = \#\{p_i > \lambda\}/[m(1 - \lambda)] := W(\lambda)/[m(1 - \lambda)]$, where $W(\lambda) = \#\{p_i > \lambda\}$. Conceptually, for a fixed λ, (8) can be estimated by

$$\frac{\hat{\pi}_0 t}{\widehat{\Pr}(P \leq t)} = \frac{W(\lambda)t}{(1 - \lambda)R(t)}. \tag{9}$$

Note that when m is large, (9) can be the estimator of both the FDR and the pFDR: Storey and Tibishrani (2003) obtain this FDR estimator by a large m approximation: FDR $\approx$ pFDR $\approx E[V(t)]/E[R(t)]$.

However, under a finite m, $R(t)$ might be 0, causing (9) undefined, and pFDR $\neq$ FDR. As such finite modification is needed for both FDR and pFDR estimators by first replacing $R(t)$ with $[R(t) \vee 1]$. This gives

$$\widehat{\text{FDR}}_\lambda(t) = \frac{\hat{\pi}_0 t}{\widehat{\Pr}(P \leq t)} = \frac{W(\lambda)t}{(1 - \lambda)[R(t) \vee 1)]}. \tag{10}$$

By the facts that pFDR = FDR / $\Pr(R(t) > 0)$ and that they wish to conservatively estimate pFDR, an estimate of pFDR can be acquired by (10) divided by a lower bound of $\Pr(R(t) > 0)$, $1 - (1 - t)^m$,

$$\widehat{\text{pFDR}}_\lambda(t) = \frac{\hat{\pi}_0 t}{\widehat{\Pr}(P \leq t)[1 - (1 - t)^m]} = \frac{W(\lambda)t}{(1 - \lambda)[R(t) \vee 1)][1 - (1 - t)^m]}. \tag{11}$$

They also suggested using nonparametric bootstrap to compute a confidence interval of the estimators. If (10) or (11) exceeds 1, then force the value to be 1. This truncation leads to smaller MSEs for the estimators [Theorem 3, Storey (2002)]. Theorem 2 in their work proves the conservative properties of $\widehat{\text{pFDR}}_\lambda(t)$ and $\widehat{\text{FDR}}_\lambda(t)$, i.e., both of them are "upward" biased [Theorem 6 of Storey, Taylor and Siegmund (2004) also provided stronger results when m is large]. Theorem 4 shows a stronger asymptotic result: $\widehat{\text{pFDR}}_\lambda(t)$ converges to pFDR(t) almost surely.

For the choice between the two estimator above, the authors recommended applying $\widehat{\text{pFDR}}_\lambda(t)$ for subsequent analyses since for t small enough, the two estimators behave differently. Namely, $\lim_{t\to 0}\widehat{\text{pFDR}}_\lambda(t) = \hat{\pi}_0(\lambda)$ and $\lim_{t\to 0}\widehat{\text{FDR}}_\lambda(t) = 0$. They argued that the latter approaches to 0 is purely driven by the term $\Pr(R(t) > 0) \to 0$, which says nothing to false discoveries.

The tuning parameter $\lambda \in [0, 1]$ indicates a bias-variance trade off in π_0: as $\lambda \to 1$, the bias becomes smaller but the variance increases; as $\lambda \to 1$ the variance becomes smaller but the bias becomes larger. A straight forward data-driven method to determine λ is to choose λ to minimize the MSE of estimator of FDR or pFDR. Algorithm 3 in Storey (2002) minimizes a bootstrap version of the MSE. Another method for choosing λ by fitting a natural cubic spline of $\hat{\pi}_0(\lambda)$ on λ is given in remark B, Storey and Tibshirani (2003). Both methods essentially compromise between variance and bias. A naive choice is $\lambda = 0.5$.

As a byproduct of (10) and (11), the estimated null probability $\pi_0(\lambda)$ has interesting properties and an interpretation in analysis. By conservatively choosing λ, i.e., choose λ so that the count $\#\{p_i > \lambda\}$ includes not only nulls but also non-nulls, one has $E[\hat{\pi}_0(\lambda)] \geq \pi_0$. Storey and Tibshirani (2003) exploit the estimate. Define $\hat{\pi}_1(\lambda) := 1 - \hat{\pi}_0(\lambda)$. Clearly, $E[\hat{\pi}_1(\lambda)] = 1 - E[\hat{\pi}_0(\lambda)] \leq \pi_1(\lambda)$, $\hat{\pi}_1(\lambda)$ provides an "average" lower bound of $\pi_1(\lambda)$. It can be interpreted as follows: suppose we have $m = 1000$ and $\hat{\pi}_1(\lambda) = 0.4$, then we can say that among 1000 tests, at least $1000 \times 0.4 = 400$ of them are "expected" to be non-nulls. The next question is which tests should we claim significant? To answer that question the q-value is established from (10) and (11) to construct MCPs.

3.2 MCP using the q-value

The q-value is an Bayesian analogy of p-value under the Bayesian two-class model. As the definition of p-value, the definition of the q-value can be defined in a general way for nested rejection regions Γ_α. Let $\alpha_i := \Pr(T_i \in \Gamma_{\alpha_i} | H_i = 0)$, the size of rejecting null hypothesis if $T_i \in \Gamma_{\alpha_i}$. If for $\alpha_1 \leq \alpha_2$, one has $\Gamma_{\alpha_1} \subseteq \Gamma_{\alpha_2}$, then Γ_α is called nested. For such regions, the q-value is defined as [Definition 2, Storey (2003)]

$$q\text{-value}(t_i) = \inf_{\{\Gamma_\alpha:\ t_i \in \Gamma_\alpha\}} \text{pFDR}(\Gamma_\alpha), \tag{12}$$

where t_i is the observed test statistic of T_i. When the rejection region is nested, the q-value(t_i) is defined to be the minimum value of pFDR if we call the feature i with observed test statistic t_i significant. The q-value can also be defined through p-value [Theorem 2, Storey (2003)]. Obviously, $\{T_i \in \Gamma\} = \{P_i \leq t\}$ implies

$$q\text{-value}(t_i) = q\text{-value}(\{p_i \leq t\}) := q\text{-value}(p_i) = \inf_{\{t:\ p_i \leq t\}} \text{pFDR}(t), \tag{13}$$

for some threshold $t \in [0, 1]$. Note that the observed statistic t_i is not t, the fix threshold for individual tests. By (13), an estimate of q-value(p_i) is $\hat{q}(p_i) := \widehat{\text{pFDR}}(t)$ with $\widehat{\text{pFDR}}(t)$ being (11). Theorem 7 in Storey, Taylor and Siegmund (2004) contends that under proper convergence conditions, $\hat{q}(p_i)$ is conservative for any $p_i \geq 0$ almost surely, i.e, for any $\delta > 0$

$\lim_{m\to\infty} \inf_{p_i \geq \delta}\{\hat{q}(p_i) - q\text{-value}(p_i)\} \geq 0$ almost surely. This conservative property is desired since we do not want to be too optimistic about a feature being significant.

Then an MCP can be derived by using $\hat{q}(p_i)$. We present the one used in Storey and Tibishirani (2003). In a large scale test where m is large, the finite modifications in (10) and (11) are not necessary. In addition, when m is large, these two quantities agree and become

$$\widehat{\text{pFDR}}_\lambda(t) = \widehat{\text{FDR}}_\lambda(t) = \frac{W(\lambda)t}{(1-\lambda)R(t)}, \tag{14}$$

which is the FDR estimate used in the article and was derived from the relationship FDR $\approx$ pFDR $\approx E[V(t)]/E[R(t)]$ for a large m; hence the independence of P_i's is not required [Recall that the independence is needed in the finite m case, Storey (2002)]. In this context, an estimate of the q-value is

$$\hat{q}(p_i) = \inf_{t \geq p_i} \widehat{\text{FDR}}_\lambda(t) = \inf_{t \geq p_i} \frac{W(\lambda)t}{(1-\lambda)R(t)} = \inf_{t \geq p_i} \frac{\hat{\pi}_0(\lambda)mt}{\#\{p_i \leq t\}}, \tag{15}$$

which is asymptotically conservative under mild convergence conditions i.e., weak dependence. An algorithm computing $\hat{q}(p_i)$ is given in remark B in Storey and Tibishirani (2003). Note that alternatively λ can be obtained by minimizing a bootstrap version of MSE of $\hat{\pi}_0(\lambda)$ [Algorithm 9, Storey (2002)].

Once we have $\hat{q}(p_i)$, we can conduct an MCP using this value. $\hat{q}(p_i)$ provides "local" information in terms of "global" criteria. By definition, the q-value can be translated as the minimum value of pFDR or FDR we will have if we call the features with p-values smaller than p_i significant. Hence an MCP is simply thresholding the q-values of tests at a level α, which is not necessarily be determined beforehand.

About the MCP, here are some additional remarks:

1. A single cutoff is not always required. One can report all the q-values of the features then decide a cutoff, which gives flexibility to the MCP. Fig 2 in Storey and Tibishirani (2003) can serve as a tool to help researchers determine a cutoff. For instance from Fig 2 (c), the researcher may decide to tolerate a slightly higher error rate to obtain significantly more discoveries. This decision may be desired for an exploratory purpose, e.g., find as many differently expressed genes as possible in an exploratory stage of a gene study.

2. The MCP that rejects individual null hypotheses if $\hat{q}(p_i) \leq \alpha$ yields FDR α as $m \to \infty$, i.e., the MCP controls the FDR in a large scale test at level α [Remark D, Storey and Tibishirani (2003)].

3.3 Relationship between Estimation and Algorithm

Despite that MCPs can be derived from two seemly quite different perspectives, Storey, Taylor and Siegmund (2004) showed that the MCPs derived via these two perspectives are actually equivalent. Consider an MCP that rejects $H_{(i)}$ with corresponding ordered observed p-values $p_{(1)} \leq ... \leq p_{(m)}$, $i = 1, ..., l$, where $l = \max\{i : \widehat{\text{FDR}}(p_{(i)}) \leq \alpha\}$. Storey (2002) proved that the BH procedure is the above MCP with $\hat{\pi}_0(\lambda) = 1$, making the BH procedure too conservative; whereas the MCP proposed by using $\widehat{\text{FDR}}(p_{(i)})$ has $\hat{\pi}_0(\lambda) \leq 1$, which will lead to more rejections while controlling the error rate at the same level. As a result, the MCP thresholding $\widehat{\text{FDR}}(t)$ in (10) is more powerful than the BH procedure.

Going further, Storey, Taylor and Siegmund (2004) presented more general results. Define an MCP that thresholds P_i's at

$$t_\alpha(\widehat{\text{FDR}}_\lambda) := \sup\{t \in [0,1] : \ \widehat{\text{FDR}}_\lambda(t) \leq \alpha\}, \tag{16}$$

where $\widehat{\text{FDR}_\lambda}(t)$ is defined as in (10). This gives an MCP developed from estimators. The BH procedure controlling the FDR at α corresponds (16) with $\lambda = 0$ or, equivalently, $\hat{\pi}_0(\lambda) = 1$. Under proper convergence conditions, the FDR of an MCP thresholding p-values at $t_\alpha(\widehat{\text{FDR}_\lambda})$ is less than or equal to α when m is large [Theorem 4, Storey, Taylor and Siegmund (2004)]. As a final remark, an MCP derived by (16) will be problematic if we choose $\alpha < \inf_{t\in[0,1]} \widehat{\text{FDR}_\lambda}(t)$ then the threshold will drop to 0 and thus the MCP rejects nothing, potentially leading to a low power. Zhang, Fan and Yu (2011) called this phenomenon "lack of identification," and proposed a method exploiting spatial information of p-values in some neighborhood of p_i to alleviate this problem.

4 Local False Discovery Rate (fdr)

4.1 Motivation for fdr

As the q-value defined previously, the local FDR (fdr) aims to provide local information for individual tests as well, albeit from a different viewpoint. Again, suppose that our tests come from a Bayesian two-group model as those in Storey (2003), i.e., H_i comes from null with probability π_0; comes from non-null with probability $\pi_1 := 1-\pi_0$. Test statistics Z_i, $i = 1, ..., m$ obtained from data through a transformation (19), satisfy

$$Z_i \sim f_0(z) \text{ if } Z_i \text{ comes from null and } Z_i \sim f_1(z) \text{ if } Z_i \text{ comes from non-null,}$$

i.e., f_0 is the density of Z_i under null, whereas f_1 is the density of Z_i under non-null. When m is large, the previously defined FDR (1) with a fixed rejection region can be approximated as a posterior [Theorem 4, Storey (2003)]. Without loss of generality, $Z_i \geq z$ or $|Z_i| \geq z$, etc, we can thus define

$$\text{FDR}(z) := \Pr(H_i \text{ is null} | Z_i \leq z) = \text{the } q\text{-value}, \tag{17}$$

which is conditioned on a rejection region $Z_i \leq z$ for a large m. The interpretation is: if we reject tests with $Z_i \leq z$ with $z \in [0, 1]$ being a fixed threshold, then the probability of false rejection for a test i is FDR(z). But such a statement ignores the fact that the probabilities of false rejection for Z_i, which is far from z, and Z_j, which is close to z, should be different. Hence the method implicitly assume "exchangeability" of the tests [Efron (2007)]. This drawback gives the motivation for defining an error rate conditioned on $Z_i = z$. Efron, Tibshirani, Storey and Tusher (2001) defined the local false discovery rate (fdr) as

$$\text{fdr}(z) := \Pr(H_i \text{ is null} | Z_i = z) = \frac{\pi_0 f_0(z)}{f(z)} := \frac{f_0^+(z)}{f(z)}, \tag{18}$$

where $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ is the mixture density. This error rate, they argued, is more suitable for empirical Bayes methods and makes more sense in the Bayesian framework since it conditions on a certain value not a region as in (17). They also proposed using logistic regression to estimate f_0/f and applying some upper bound of π_0 to estimate it. Note that typically, fdr $>$ FDR so they suggest that the use the FDR (q-value) is often too optimistic. Literature suggests thresholding fdr at 0.2, corresponding to thresholding the FDR (q-value) between 0.05 and 0.15.

Here, we mainly focus on introducing a more general method in Efron (2007). We will concentrate on estimation but will briefly introduce power calculation and accuracy analysis of fdr. Note that these estimating methods do not require independence of Z_i's but need to have a large m and that the closed-form power and accuracy analysis given by this fdr method are not in the FDR methods.

4.2 Preliminaries

As in literature, we use gene expression level tests as an example. Suppose that we would like to test the differences of gene expression levels for each gene i between a normal cell and a cancered cell, i.e., the null hypothesis $H_{0i} : \mu_{X_i} - \mu_{Y_i} = 0$ with X_i from normal cell microarrays, Y_i from cancered cell microarrays, and μ is the mean of gene expression levels among normal or cancered microarrays for gene i, $i = 1, ..., m$. Assume we have l microarrays from normal cells and k of those from cancered cells. Here, they make a z score transformation:

$$Z_i = \Phi^{-1}(F_{l+k-2}(T_i)), \tag{19}$$

where Φ is the c.d.f. of standard normal distribution, F is the c.d.f. of Student's t distribution with degrees of freedom $(l + k - 2)$, and T_i is the two-sample t statistic. Ideally, if the gene expression levels distribute as normal and are independent, Z_i, the test statistics used in analysis, should obey $Z_i \sim N(0, 1)$ under null. In the context of section 4.1, we have $N(0, 1) \overset{d}{=} f_0(z)$ in an ideal situation. Efron (2007) called $N(0, 1) \overset{d}{=} f_0(z)$ the theoretical null.

However, in real world, the theoretical null might not hold for some reasons. Plotting the histogram of Z_i's and density of $N(0, 1)$ can provide a check as in Fig 1 of Efron (2007). If the histogram of Z_i's shows a clear deviation from the standard normality, then one has to estimate $f_0(z)$ since the misspecification of null density will undermine the inference. The estimated null, $\hat{f}_0(z)$, is called the empirical null. Several situations can cause the violation of the theoretical null, such as non-normality of gene expression levels, the existence of unobserved covariates, correlation between $(l+k)$ arrays for gene i, or correlation across genes. Efron (2007) and Efron (2012) give a thorough discussion of the effects and corresponding remedies, e.g., the normality problem can be avoided via permutation analysis. The price of using the empirical null is the increase in variability of $\hat{\text{fdr}}$ [Table 4, Efron (2007)]. Nonetheless, if the data apparently does not follow $N(0, 1)$, such price must be paid.

4.3 Estimation of fdr

Here, the estimation of fdr in Efron (2007) is presented with some supplements from Efron (2012). The methods below can be implemented by the R package *locfdr*. They estimated fdr by

$$\hat{\text{fdr}}(z) = \frac{\hat{\pi_0}\hat{f}_0(z)}{\hat{f}(z)} = \frac{\hat{f_0^+}(z)}{\hat{f}(z)}, \tag{20}$$

which can utilize the fact that f_0^+ and f are smooth, see section 7.4 Efron (2012). Even under non-null hypothesis, Z_i's approximately distribute as $N(\mu_i, \sigma_i^2)$ as shown in Fig 7.6 from Efron (2012).

4.3.1 Estimation of $f(z)$

As mentioned previously, an approximation of Z_i to normality gives a good reason to believe that $f(z)$ is smooth. Efron (2007) recommended using a Poisson generalized linear model (GLM) to model discretized Z_i's for that the method leads to closed-form analysis of accuracy and easy implementation. The method is called Lindsey's method. Section 5.2 of Efron (2012) gives an illustration of Lindsey's method, which is summarized in the following steps:

1. Partition the range $\mathcal{Z}$ of Z_i's into K bins $\mathcal{Z}_k$ with equal width d such that $\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k$.

2. Let $y_k = \#\{Z_i \in \mathcal{Z}_k\}$ be independent Poisson counts with rate ν_k, i.e., $y_k \overset{ind}{\sim} Poisson(\nu_k)$ for all k and x_k be the center point of $\mathcal{Z}_k$.

3. Then the expected counts of y_k is approximately $\nu_k = mdf(x_k)$.

4. The standard Poisson GLM fits $\log \nu_k$ by a linear model, equivalently,

$$\log f(x_k) = \sum_{j=0}^{J} \beta_j x_k^j,$$

with $\hat{\beta}_0$ chosen to let $f(x_k)$ integrate to 1 and $\hat{\beta}_j$'s solved by the maximum likelihood method. Thus one obtains $\hat{f}(z)$. $\hat{\nu_k} = md\hat{f}(x_k)$ is the smoothed version of y_k, which is more stable than y_k.

The dependence of Z_i's will cause over/under-dispersion of y_k, but it has little impact on ν_k. Hence they use $\hat{\nu_k} = md\hat{f}(x_k)$ in fdr. Alternatively, one can also use a natural spline with degree J to fit $\log f$. $J = 7$ is the default in *locfdr*.

4.3.2 Estimation of $f_0^+(z)$ (Empirical Null)

The needs of the empirical null are justified above. Efron (2007) proposed two different strategies to estimate $f_0^+(z)$: central matching and MLE fitting. Both these methods rely on the "zero assumption," which assumes that most Z_i's near 0 are truly null cases. The $p-$value version of this assumption is that most $p-$values near 1 are truly null cases. The estimator proposed in Storey (2002) in section 3.1.2 also utilized this assumption with the tuning parameter λ reflects the zero assumption region, indicating a bias and variance trade-off in $\hat{\pi_0}$. Some form of zero assumption must be incorporated due to the absence of parametric assumption of f_1 and it is unidentifiable [Efron (2007))]. In FDR literature, the distribution of non-null case is rather intractable than that of null [Yekutieli and Benjamini (1999)]. Nonetheless, the zero assumption is more believable when $\pi_0 \geq 0.9$, i.e., the multiple test is sparse, which is often in high-dimensional applications, such as gene expression testing. Furthermore, under sparsity $\pi_0 \geq 0.9$, it can be shown that central matching is nearly unbiased [Efron (2004)]. Both methods assume that $f_0(z) \sim N(\delta_0, \sigma_0^2)$, which, as argued in Efron (2012), is usually the case and justified by an distributional approximation.

Central matching exploits $\hat{f}(z)$ obtained in Lindsey's method. By the normality and zero assumptions, a quadratic function of $\log \hat{f_0^+}(z)$ is first fit to $\log \hat{f}(z)$ near $z = 0$, e.g., using least square estimation to solve coefficients to the central one third part of Z_i's. Namely, one models $\log \hat{f_0^+}(z)$ as

$$\log \hat{f_0^+}(z) = \hat{\beta}_0 + \hat{\beta}_1 z + \hat{\beta}_2 z^2.$$

Then use normality assumption again to solve for $\{\hat{\pi_0}, \hat{\delta_0}, \hat{\sigma_0}^2\}$ by $\{\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2\}$ obtained from the above model, i.e., solve $\log \hat{f_0^+}(z) = \hat{\beta}_0 + \hat{\beta}_1 z + \hat{\beta}_2 z^2 = \log \hat{\pi_0} + \log \varphi_{\hat{\delta_0}, \hat{\sigma_0}^2}$, where $\varphi_{\delta_0, \sigma_0^2}$ is the $N(\delta_0, \sigma_0^2)$ density. See Fig 5 in Efron (2007).

MLE fitting does not need $\hat{f}(z)$ estimated from Lindsey's method; hence it does not rely on the discretization of range $\mathcal{Z}$. The zero assumption is needed however. Suppose that a predetermined zero region $\mathcal{A}_0$ is given, e.g., $[-x_0, x_0]$ for some $x_0 \in \mathbb{R}_+$ [Efron (2007)]. Then the zero assumption implies $f_1(z) = 0$ if $z \in \mathcal{A}_0$. As such the Z_i's in $\mathcal{A}_0$ are all nulls and follow $N(\delta_0, \sigma_0^2)$ by normality assumption. As a result, the likelihood of $\{\pi_0, \delta_0, \sigma_0^2\}$ follows some kind of truncated normal distribution given in equation (4.12) of Efron (2007). Then by the standard MLE procedure and invariance property of the MLEs, the MLE $\{\hat{\pi_0}, \hat{\delta_0}, \hat{\sigma_0}^2\}$ can be acquired.

Both methods have merits and drawbacks. Central matching enables us to see how far the empirical null deviates from $N(0, 1)$. If the test is sparse enough, $\pi_0 \geq 0.9$, then $\{\hat{\delta_0}, \hat{\sigma_0}^2\}$

solved by central matching is nearly unbiased. However, central matching gives estimates that are more numerically variable than those acquired from MLE fitting and are sensitive to $\mathcal{Z}$. MLE fitting is more stable but depends heavily on the choice of the zero region $\mathcal{A}_0$. [Table 3, Efron (2007)].

4.3.3 Putting Things Altogether

Upon completing the above procedures, we are ready to estimate fdr by

$$\hat{\text{fdr}}(x_k) = \frac{md\hat{f}(x_k)}{m\int_{\mathcal{Z}_k} \hat{f_0^+}(z)\, dz}, \tag{21}$$

for k-th bin. If we claim the cases within this bin significant, then about $\hat{\text{fdr}}(x_k)$ of them will be false discoveries. An MCP based on fdr can therefore acquired, for instance, calling the cases in the k-th bin significant if $\hat{\text{fdr}}(x_k) \leq 0.2$.

4.4 Power Calculation and Accuracy Analysis

The approaches from Efron (2007) also allow power calculation and accuracy analysis for the estimated fdr. The power of a test is defined to be the probability of rejecting truly non-null cases. With the estimates obtained from (21), they developed three useful tools to access power. First, $\widehat{E[\text{fdr}]_1}$, the expectation of $\hat{\text{fdr}}$ with respect to $\hat{f}_1$. Loosely speaking, this quantity gives the value of fdr under test statistics come from non-nulls. Hence a small value of $\widehat{E[\text{fdr}]_1}$, e.g., 0.2 indicates good power. This estimate can be obtained by equation (3.8) in Efron (2007). Additionally, they derived the effect of increasing Z_i by a factor of c for gene i on $\widehat{E[\text{fdr}]_1}$. This answers the question: if we were to increase the number of subjects, how would the gain of power become? Second, the non-null counts $\hat{y_{1k}} := [1 - \hat{\text{fdr}}(x_k)]y_k$ as shown in Fig 1 and Fig 2 in the paper. $\hat{y_{1k}}$, by definition, has the interpretation: the expected counts of truly non-null cases in the k-th bin. In practice, the more the value of $\hat{y_{1k}}$'s in the rejection area, e.g, bins having $\hat{\text{fdr}}(x_k) \leq 0.2$, the more power a study has, i.e., among cases reported significant many of them are truly non-null. Third, the empirical non-null c.d.f. defined as

$$\hat{G}_1(t) = \frac{\sum_{k\in\mathcal{K}} \hat{y_{1k}}}{\sum_{k=1}^{K} \hat{y_{1k}}},$$

where $\mathcal{K} = \{k : \hat{\text{fdr}}(x_k) \leq t\}$. Given the same t, the study has greater $\hat{G}_1(t)$ possesses more power.

Finally, the influence of count y_k on the estimation of fdr is analyzed. The influence function $(\partial \log(\text{fdr})_k/\partial y_l)_{K\times K}$ has closed-forms when using central matching and MLE fitting [Lemma 1 and Lemma 2, respectively, Efron (2007)]. Then the closed-forms of covariance under these two approaches follow [Equation (5.24) and (5.26), Efron (2007)].

4.5 Compare the q-value and fdr

Fig 3 in Efron (2007) shows the geometric relationship of FDR and fdr. Analytically, we have $\text{FDR}(z) = E_f[\text{fdr}(Z)|Z \leq z]$, i.e., FDR is the average of fdr conditioned on $Z \leq z$. The expectation E_f means taking expectation with respect to density f. Both the q-value, which is the FDR, by definition (17) under m large enough, and fdr provide significance of an individual test. In general fdr > FDR, implying that FDR is often too optimistic. They both have merits and drawbacks. In short, the q-value ignores the impact of different values of test statistics on

individual significance. It simply presumes that tests are exchangeable in some sense, whereas fdr takes such effects into account. The q-value measures individual significance in terms of FDR of a multiple test, incorporating multiplicity. The fdr, in contrast, does not consider the multiplicity problem. Ideally, both criteria should be considered in an analysis to have better inference, e.g., Kilne, Rose and Walters (2022) reported both of them in their multiple test.

5 Dependence

In the above sections, we mainly focus on approaches that rely on the independence of test statistics or p-values. In section 2.1 the BH procedure requires such an independence premise, albeit it can be relaxed to be only independent null p-values. The adaptive method in section 2.2 has theoretical FDR control when the p-values are independent. When the p-values have mild correlations, the approach shows FDR control in the numerical experiments. The estimators in section 3 also depend on such an assumption except for the cases when m is large. Some good properties such as conservativeness and Bayesian posterior interpretation still hold for a large m under weak dependence [Convergence of empirical null probability, empirical power, and empirical size, Theorem 4, Storey (2003)]. On the contrary, the local FDR approach in section 4 bypasses the independence restriction but demands a large m instead. Dependence may cause the null density deviates the theoretical null $N(0,1)$ nonetheless.

In addition to the local FDR approach, there are other research focusing on dependence, which researchers might encounter in some studies. In many applications, the test statistics are often positively correlated. When such a dependent relationship, it is possible that modifying original MCPs will increase the power.

Inspired by Westfall and Young's resampling p-value MCP controlling the FWER, Yekutieli and Benjamini proposed a resampling p-value MCP that controls the FDR when test statistics are highly correlated in 1999. Conceptually, if the joint behavior of $\mathbf{P} = \{P_1, ..., P_m\} := \{\mathbf{P}_0, \mathbf{P_1}\}$ is known, where $\mathbf{P}_0$ and $\mathbf{P}_1$ are sets of p-value of null and non-null, respectively, then the behavior of the FDR is known. Since $\mathbf{P}_1$ is intractable, they first conditioned on $S = s$, then using resampling scheme to approximate the distribution of $\mathbf{P}_0$ to estimate the FDR conditioned on $S = s$ for fixed rejection regions similar to Storey (2002). The resulting estimate can hence be converted to an MCP under dependence [Section 5, Yekutieli and Benjamini (1999)]. Their simulation shows a greater power over the BH procedure when high correlation emerges.

Yekutieli and Benjamini (2001) proved general versions of the BH procedure under some forms of positive correlation of test statistics. The forms of positive correlation they resorted to can be classified as a special case of positive regression dependency on each one from a subset of the set of null indices I_0, or PRDS on I_0 for short. They showed in Theorem 1.2 of their article that when the joint distribution of the test statistics T_i's is PRDS on I_0, the BH procedure still controls the FDR under $(m_0/m)\alpha$. Many cases can be identified as PRDS on I_0, for example, when T_i's are jointly multivariate normal distribution with covariances between any two T_i and T_j, $i \neq j$ and $i, j \in I_0$ are non-negative; $|\mathbf{T}|/S$, where $\mathbf{T} = (T_1, ..., T_m)$, $|\mathbf{T}|$ is PRDS on I_0, and $S > 0$ is a chi-square estimated estimator, is also PRDS on I_0 [Case 1 and Case 3, respectively, Yekutieli and Benjamini (2001)]. Theorem 1.3 further proved a universal threshold for an MCP under any unknown dependence structure, but the MCP is somewhat conservative.

Lastly, in addition to developing MCPs combating dependent test statistics, one may wish to analyze the effect of dependence on the FDR estimates, say, those in section 2. Schwartzman and Lin (2011) conducted such analyses by approximating mean, variance, distribution, and quantiles of the FDR estimator (10) for cases whose test statistics are marginally distributed as normal or chi-square under arbitrary dependence structure. They evaluated the mean and variance of the FDR estimator in several stages. First, a critical component that determines the distributional behaviors of (10) is $R(t) = \#\{p_i \leq t\}$ for all i which can be modeled by a

binomial-beta mixture and then approximate the binomial-beta mixture model by a gamma-Poisson mixture model, resulting a marginal distribution of $R(t)$, a negative binomial with a mean parameter and an overdispersion parameter determining the variance. Second, under distributional assumptions on test statistics, it can be shown that the variance terms consisting overdispersion parts can be expanded as polynomials [Theorem 2 and Theorem 3, Schwartzman and Lin (2011)]. In the final stage, use method of moments to solve these parameters then apply Theorem 4 to obtain estimates of mean, variance, distribution, and quantiles of the FDR estimator (10). This allows one to evaluate the performance of the FDR estimator under general dependence in a more comprehensive way in a z or χ^2 test.

6 Applications

MCPs controlling false discovery rates have been widely used in applied fields such as gene studies, multiple endpoints tests, functional Magnetic Resonance Imaging, and among others. These studies share a common feature: a large m and, typically a large π_0, i.e., a sparse test in high dimensions. The FDR methods gain advantages over the FWER methods due to its scalable nature. Many applications consider the FDR MCPs exploratory data analysis tools to find out as many as possible discoveries, and then conduct further investigations. Conventional work often use the above-mentioned field as examples of FDR methodology.

An interesting application of the q-value method and local FDR in economics is given in Kilne, Rose and Walters (2022) in which they empirically verified whether there exists systematic discrimination among 108 large U.S. companies. Particularly they focused the presence of race and gender discrimination. They defined the systematic racial discrimination between white and Black applicants as $\Delta_i :=$ contact rates of white applicants $-$ contact rates of Black applicants for company i. In section X, they investigated individual estimates and tested $H_{0i} : \Delta_i = 0$ with alternative H_{1i} are one-tailed or two-tailed for $i = 1, ..., 108$. The later multiple test is where the q-value and the local FDR can be utilized. Table VIII in their work shows the estimation results where the q-value is estimated by (15), where the tuning parameter λ obtained by Storey (2002) and Storey and Tibishirani (2003) are both reported.

They argued that the use of FDR control MCPs allows one to specify which companies have systematic discrimination while controlling the false discoveries [numbers of companies that are called discriminatory but they are actually not] under a certain level. This provides regulators insights about which companies they should audit. Reporting these results to those companies also confers them an opportunity to improve their recruiting process.

7 Simulation Studies

This part is still underway and will be uploaded to my github.

8 References

1. Benjamini, Y. & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289-300.

2. Benjamini, Y. & Yekutieli, D. (2001). The Control of The False Discovery Rate in Multiple Testing under Dependency, *Annals of Statistics,* **29**, 1165–1188.

3. Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis, *Journal of the American Statistical Association*, **99**, 96–104.

4. Efron. B. (2007). Size, Power and False Discovery Rates. *Annals of Statistics,* **35(4)**, 1351 – 1377.

5. Efron B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, Vol. 1.* Cambridge, UK: Cambridge Univ. Press.

6. Efron, B., Tibshirani, R., Storey, J. D, & Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association*, **96(456)**, 1151-1160.

7. Food and Drug Administration. (2022). *Multiple Endpoints in Clinical Trials: Guidance for Industry.* Maryland, United States of America: Author.

8. Kline, P., Rose. E. K., & Walters. C. R. (2022). Systemic Discrimination Among Large U.S. Employers. *The Quarterly Journal of Economics*, **137(4)**, 1963-2036.

9. Liu. F & Sarkar. S. K. (2011). A New Adaptive Method to Control the False Discovery Rate. *Recent Advances in Biostatistics*, **4**, 3-26.

10. Schwartzman, A. & Lin, X. (2011). The Effect of Correlation in False Discovery Rate Estimation. *Biometrika*, **98(1)**, 199–214.

11. Storey. J. D. (2002). A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **64(3)**, 479-498.

12. Storey, J. D. (2003). The Positive False Discovery Rate: A Bayesian Interpretation and the q-value. *Annals of Statistics,* **31**, 2013–2035.

13. Storey, J. D., Taylor, J. E. & Siegmund, D. (2004). Strong Control, Conservative Point Estimation, and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society. Series B (Statistical Methodology),* **66**, 187–205.

14. Storey, J. D. & Tibshirani, R. (2003). Statistical Significance for Genome-Wide Studies, *Proceedings of the National Academy of Sciences,* **100**, 9440–9445.

15. Storey. J. D., Bass. A. J., Dabney. A., & Robinson. D. (2015). qvalue: Q-value Estimation for False Discovery Rate Control. *R Package*, https://github.com/StoreyLab/qvalue.

16. Yekutieli, D. & Benjamini, Y. (1999). Resampling-Based False Discovery Rate Controlling Multiple Test procedures for correlated test statistics. *Journal of Statistical Planning and Inference,* **82(1–2)**, 171-196.

17. Zhang, C., Fan, J., & Yu, T. (2011). Multiple Testing via FDRL for Large-Scale Imaging Data. *Annals of Statistics,* **39(1)**, 613 – 642.