On Storey's direct approach to false discovery rates

Marko Lalović
July 2019

Abstract

Storey [1] shed a new light on multiple-testing problem with the definition of positive false discovery rate measure and gave new perspective with direct approach to false discovery rates.

We provide formal introduction to multiple-testing problem and intuitive explanation of Storey's direct approach. We demonstrate the duality between Storey's direct approach and Benjamini-Hochberg (BH) procedure [2] when using Storey's estimate $\widehat{\pi}_0$ of the proportion of null hypotheses π_0 in BH procedure.

We show that the estimator $\widehat{\pi}_0$ of the proportion of null hypotheses π_0 plays a key role in multiple-testing problem and that the approach we take doesn't matter much. We confirm this using simulations where performances were practically identical for both methods.

We show that Storey's estimator $\widehat{\pi}_0$ can be very upward biased when distance between distributions under null hypothesis and alternative hypotheses is small. In our simulations this upward bias of $\widehat{\pi}_0$ was reduced if we increased the value of the tuning parameter λ and especially by tuning this λ parameter using the bootstrap method. However this bootstrap method can result in underestimation of π_0, as already explained in [3].

In case of dependence Storey suggested the same approach [4]. We show, this time only using simulation, that dependence can lead to a very high overestimation of false discovery rate.

Keywords: Multiple testing, Multiple comparisons, False discovery rate

Contents

1 Introduction

1.1 Single-hypothesis testing

Following the scientific method, researchers normally want to establish the truth of a statement by showing that the opposite appears to be false. In general, a *hypothesis* is a proposed explanation for a phenomenon that one can test. In statistics, the hypotheses involve a parameter θ whose value is unknown but must lie in a certain parameter space Ω. We assume that Ω can be partitioned into two disjoint subsets Ω_0 and Ω_1 and define two hypotheses:

$$H_0\colon \theta \in \Omega_0,$$
$$H_1\colon \theta \in \Omega_1.$$

The hypothesis H_1 is the statement that confirms the theory and it is called an *alternative hypothesis*. The opposite H_0 is called a *null hypothesis*. We must decide which of the hypotheses appears to be true. A procedure for deciding this is called a *test procedure* or simply a *test*. Before we have to decide which hypothesis to choose, we assume that we can observe a random sample $\mathbf{X} = (X_1, ..., X_n)$ drawn from a distribution that involves the unknown parameter θ. Let $T = r(\mathbf{X})$ be a statistic and let Γ be a subset of the real line and suppose that the test procedure is of the form: reject H_0 if $T \in \Gamma$. Then T is called a *test statistic* and Γ a *rejection region*. A problem of this type is called a *single-hypothesis testing*.

In single-hypothesis testing we only have two kinds of errors we might make. A *type I error* occurs when we reject the null H_0 when in fact H_0 is true. A *type II error* occurs when we fail to reject the null H_0 when in fact H_0 is false. The *significance level* α is the upper bound on the probability of type I error:

$$\Pr(T \in \Gamma \mid H_0 \text{ is in fact true }) \leq \alpha. \tag{1}$$

The *detection power* π of a test is the probability of rejecting H_0 when in fact H_0 is false. Statistical practice is to choose the value of α, say 0.01, and then find the rejection region Γ_α that maximizes the detection power and satisfies Eq. 1, that probability of type I error is at most α. The detection power π is a function of rejection region Γ_α:

$$\pi(\Gamma_\alpha) = \Pr(T \in \Gamma_\alpha \mid H_0 \text{ is in fact false }). \tag{2}$$

Typically, rejection regions Γ_α and $\Gamma_{\alpha'}$ for different values α and α', are nested in this sense:

$$\text{if} \quad \alpha < \alpha' \quad \text{then} \quad \Gamma_\alpha \subset \Gamma_{\alpha'}. \tag{3}$$

For the given observation $T = t$, the smallest significance level α at which a null hypothesis would be rejected, is called the *p-value*:

$$p(t) = \inf\{\alpha : t \in \Gamma_\alpha\}. \tag{4}$$

1.2 Multiple-hypothesis testing

Now consider simultaneous testing a collection of m hypotheses, called a *family of hypotheses* $\mathcal{F}$, where for $i = 1, ..., m$, we test H_{0i} versus H_{1i} on the basis of test statistics given in the form of p-values p_i. Let $H_1, ..., H_m$ be random variables, where

$$H_i = \begin{cases} 0 & \text{if } H_{0i} \text{ is in fact true,} \\ 1 & \text{if } H_{1i} \text{ is in fact true.} \end{cases}$$

A *multiple-testing procedure* is a decision function ϕ, which for a given p-values $p_1, ..., p_m$ assigns decision values $D_1, ..., D_m$:

$$\phi(p_i) = D_i, \quad i = 1, ..., m, \tag{5}$$

where $D_i = 1$ if hypothesis H_{0i} is rejected and $D_i = 0$ otherwise, for each $i = 1, ..., m$. Table 1 describes the possible outcomes from m hypothesis tests, where $I_0 \subset \{1, ..., m\}$ is the index of true null hypotheses, $m_0 = |I_0|$ is the number of true null hypotheses, $V = \sum_{i=1}^{m}(1 - H_i)D_i$ is the number of true null hypotheses which are rejected (type I errors), $R = \sum_{i=1}^{m} D_i$ is the number of all rejected null hypotheses, etc.

	Accepted	Rejected	Total
I_0	U	V	m_0
$\{1, ..., m\} \setminus I_0$	T	S	$m - m_0$
Total	W	R	m

Table 1: All possible outcomes from m hypothesis tests.

If the decision function ϕ is based on rejection region of the form $[0, \gamma]$ for some threshold $\gamma > 0$:

$$\phi(p_i) = \mathbb{1}(p_i \leq \gamma), \tag{6}$$

then the random variables $U, V, ..., R$ depend only on this threshold γ, e.g. $R(\gamma) = \sum_{i=1}^{m} \mathbb{1}(p_i \leq \gamma)$.

Good scientific practice requires the specification of certain type I error measure control to be done prior to the data analysis. The problem is to find a multiple-testing procedure which maximizes detection power and satisfies certain conditions involving type I error measures. A problem of this type is called a *multiple-testing problem*. It is described in [5] or in the context of assessing the feature significance in [6].

We say that some multiple-testing procedure ϕ has *strong level α control* of some error measure EM, if it satisfies the condition:

$$\text{EM}_\phi \leq \alpha \tag{7}$$

for any configuration of true and false hypotheses in the family of m hypotheses $\mathcal{F}$.

The first type I error measure that was suggested is the *family-wise error rate (FWER)*. This is the probability that we make at least one type I error among the family $\mathcal{F}$ of m hypotheses using some multiple-testing procedure ϕ:

$$\text{FWER}_\phi = \Pr(V \geq 1). \tag{8}$$

1.3 Multiple-testing procedures

1.3.1 Bonferroni correction

Simple multiple-testing procedure is the Bonferroni correction with the following decision function ϕ. Reject H_{0i}, if $p_i \leq \alpha/m$:

$$\phi(p_i) = \mathbb{1}(p_i \leq \frac{\alpha}{m}). \tag{9}$$

Proposition 1. *Bonferroni correction ensures strong level α control of FWER:*

Proof.

$$\text{FWER}_{\text{Bonf.}} = \Pr(\bigcup_{i \in I_0} \{p_i \leq \frac{\alpha}{m}\}) \leq \sum_{i \in I_0} \Pr(p_i \leq \frac{\alpha}{m}) = m_0 \cdot \frac{\alpha}{m} \leq \alpha. \tag{10}$$

□

This control comes with no distributional assumptions (e.g. independence) on the test statistics (p-values) or assumptions on the proportion of null hypotheses m_0/m. Unfortunately, Bonferroni correction has very little detection power when m is large.

Now we can try to improve detection power of this simple multiple-testing procedure by constructing an estimator $\widehat{m}_0$ of the number of true null hypotheses and simply replace m in decision function ϕ with the estimator $\widehat{m}_0$ to get a new decision function ϕ':

$$\phi'(p_i) = \mathbb{1}(p_i \leq \frac{\alpha}{\widehat{m_0}}). \tag{11}$$

If $\widehat{m}_0 < m$, we can achieve improvement in detection power using ϕ' instead of ϕ, because the set of hypotheses rejected by ϕ are contained in the set of hypotheses rejected by ϕ'.

Proposition 2. *If the estimator $\widehat{m}_0$ tends to at most overestimate m_0:*

$$E(\widehat{m}_0) \geq m_0, \tag{12}$$

then improved Bonferroni correction ϕ' satisfies strong level α control:

$$FWER_{\phi'} \leq \alpha. \tag{13}$$

Proof. Simply replace m with $\widehat{m}_0$ in the proof of Proposition 1:

$$\text{FWER}_{\phi'} \leq \frac{m_0}{\widehat{m}_0} \cdot \alpha. \tag{14}$$

Take expected value on both sides and use monotonicity of expected value and the fact that expected value of FWER is FWER:

$$\text{FWER}_{\phi'} \leq \frac{m_0}{\text{E}(\widehat{m}_0)} \cdot \alpha. \tag{15}$$

Use $\text{E}(\widehat{m}_0) \geq m_0$:

$$\text{FWER}_{\phi'} \leq \frac{m_0}{\text{E}(\widehat{m}_0)} \cdot \alpha \leq \alpha. \tag{16}$$

□

From this example of using FWER error measure and Bonferroni multiple-testing procedure we see, that improvement in detection power and control of type I error measure depends on the properties of an estimator $\widehat{m}_0$ of m_0 (or equivalently $\widehat{\pi}_0$ of the proportion $\pi_0 = m_0/m$) of the number of true null hypotheses in the family $\mathcal{F}$. More precisely, lower expected value of estimator $\widehat{m}_0$ ($\widehat{\pi}_0$) leads to a greater detection power, but it also affects type I error measure.

1.3.2 Benjamini-Hochberg procedure

Another way to improve the detection power is by using a more appropriate type I error measure. Benjamini and Hochberg [2] introduced a new type I error measure called *false discovery rate (FDR)*. This is the expected proportion of false positive results among all the rejected hypotheses when using some multiple-testing procedure ϕ:

$$\text{FDR}_\phi = \text{E}\left(\frac{V}{R} \mid R > 0\right) \cdot \Pr(R > 0). \tag{17}$$

Proposition 3. *For any multiple-testing procedure ϕ:*

$$FDR_\phi \leq FWER_\phi \tag{18}$$

with equality if all null hypotheses in $\mathcal{F}$ are true.

Proof.

$$\text{FDR}_\phi = \text{E}\left[\frac{V}{R} \cdot \mathbb{1}(R > 0)\right] \leq \text{E}\left[\mathbb{1}(V \geq 1)\right] = \text{FWER}_\phi. \tag{19}$$

When all null hypotheses in $\mathcal{F}$ are true, $V = R$ and $\text{FDR}_\phi = \text{FWER}_\phi$. □

Therefore we might achieve improvement in detection power by controlling FDR error measure instead of FWER error measure.

The following multiple-testing procedure is called *Benjamini-Hochberg (BH) procedure*. For given level α, given ordered, observed p-values $p_{(1)}, ..., p_{(m)}$, calculate:

$$\widehat{k} = \max\{k : p_{(k)} \leq \alpha \cdot k/m\}, \tag{20}$$

and reject all null hypotheses corresponding to $p_{(i)}$ for $i \leq \widehat{k}$.

For this procedure Benjamini and Hochberg [2] showed the following. If test statistics (in our case p-values) are independent and in some cases of dependence, regardless of the distribution of the test statistics (in our case distribution of p-values) when H_0 is false, this procedure has a strong level α control of FDR. Moreover it satisfies the property:

$$\text{FDR}_{\text{BH}} \leq \pi_0 \cdot \alpha \leq \alpha, \tag{21}$$

where π_0 is the proportion of true null hypotheses:

$$\pi_0 = \frac{m_0}{m}. \tag{22}$$

1.3.3 Storey's approach

Storey suggested [1] a modified version of FDR error measure called the *positive false discovery rate (pFDR)*:

$$\text{pFDR}_\phi = \text{E}\left(\frac{V}{R} \mid R > 0\right), \tag{23}$$

when using some multiple-testing procedure ϕ. It holds pFDR $\leq$ FDR:

$$\text{pFDR}_\phi \leq \text{pFDR}_\phi \cdot \Pr(R > 0) = \text{FDR}_\phi, \tag{24}$$

because $\Pr(R > 0) \leq 1$. Therefore we might achieve improvement in detection power by controlling error measure pFDR instead of FDR. In case when $\Pr(R > 0)$ is much less than 1, FDR error measure might be misleading and Storey's pFDR error measure is more appropriate. pFDR error measure also has a clean Bayesian interpretation, which we show in Chapter 2 and it is described in [7].

Storey suggested in [1] a more direct approach to multiple-testing problem. Rather than fixing level α control of some error measure and then estimating the rejection region ($\hat{k}$ in BH procedure) satisfying level α control (FDR $\leq \alpha$ for BH procedure), we instead fix the rejection region and then estimate some error measure, preferably pFDR.

Important part of Storey's direct approach is Storey's method of estimating the proportion of true null hypotheses $\pi_0 = m_0/m$. For a good estimator of (p)FDR, we require a good estimator of π_0. This problem of estimating π_0 is described in [8] and [9]. Storey's method of estimating π_0 is based on the following idea. Since the m observed p-values contain information about the proportion of true null hypotheses π_0, Storey introduces a tuning parameter $\lambda \in (0, 1)$ and use information in

$$\frac{\#\{p_i > \lambda\}}{m} \tag{25}$$

to estimate π_0.

In cases where $m_0 << m$ (or equivalently $\pi_0 << 1$), direct approach using Storey's estimators has much more detection power than BH procedure, which takes m instead of an estimate of m_0. Equivalently, we can improve detection power of BH procedure by using an estimate of m_0, but it is no longer guaranteed to achieve FDR control at the desired level. One advantage of using direct approach instead of BH procedure is that we can estimate pFDR instead of FDR.

1.4 Overview

After the formal derivation of Storey's estimators for direct approach in Section 2, we present results in Sections 3 and 4 of detection power comparison of BH procedure and direct approach using Storey's estimators. We formally justify and generalize this results. In Section 5 we show that Storey's estimator $\widehat{\pi}_0$ can be very upward biased when distance between distributions under H_0 and H_1 is small. It follows from Section 2 that the same bias is present in the estimation of type I error measure. Lastly in Section 7, we show with simulation that dependence can lead to very high overestimation of FDR.

2 Derivation of Storey's estimators for direct approach

Direct multiple-testing procedure using Storey's estimates is based on the following Theorem 4 which is valid under the following assumptions and the assumption that rejection region Γ_α satisfies the nesting property Eq. 3. Consider simultaneous testing of m null hypotheses, where for $i = 1, ..., m$, we test H_{0i} versus H_{1i} on the basis of test statistics p_i, given in the form of p-values and where significance region is $[0, \gamma]$, for a given threshold $\gamma > 0$. Assume $H_i \overset{iid}{\sim} \text{Bernoulli}(1 - \pi_0)$, where $H_i = 0$ when H_{0i} is in fact true and $H_i = 1$ when H_{0i} is in fact false. Assume p_i is uniformly distributed on $[0, 1]$ when H_{0i} is in fact true.

Denote the density of p_i, when H_{0i} is in fact false, with $h(p)$ for $p \in [0, 1]$.

Moreover, the p-values p_i are iid continuous random variables P with density:

$$f(p) = \pi_0 + (1 - \pi_0) \cdot h(p), \qquad p \in [0, 1]. \tag{26}$$

Theorem 4. *The posterior probability for a given implicit prior probability π_0 is:*

$$\textit{pFDR}(\gamma) = \textit{Pr}(H = 0 \mid P \leq \gamma). \tag{27}$$

Proof.

$$\text{pFDR}(\gamma) = \text{E}\left[\frac{V(\gamma)}{R(\gamma)} \mid R(\gamma) > 0\right] \tag{28}$$

$$= \sum_{k=1}^{m} \text{E}\left[\frac{V(\gamma)}{R(\gamma)} \mid R(\gamma) = k\right] \cdot \text{Pr}(R(\gamma) = k \mid R(\gamma) > 0) \tag{29}$$

$$= \sum_{k=1}^{m} \frac{1}{k} \cdot \text{E}\left[V(\gamma) \mid R(\gamma) = k\right] \cdot \text{Pr}(R(\gamma) = k \mid R(\gamma) > 0) \tag{30}$$

Given $R(\gamma) = k$ and assumption of independent p-values, it follows that $V(\gamma)$ is a binomial random variable, with k trials and probability of success $\text{Pr}(H = 0 \mid P \leq \gamma)$ with expected value:

$$\text{E}\left[V(\gamma) \mid R(\gamma) = k\right] = k \cdot \text{Pr}(H = 0 \mid P \leq \gamma). \tag{31}$$

Combining 30 and 31 we get:

$$\text{pFDR}(\gamma) = \sum_{k=1}^{m} \frac{1}{k} \cdot k \cdot \text{Pr}(H = 0 \mid P \leq \gamma) \cdot \text{Pr}(R(\gamma) = k \mid R(\gamma) > 0) \tag{32}$$

$$= \text{Pr}(H = 0 \mid P \leq \gamma) \cdot \sum_{k=1}^{m} \text{Pr}(R(\gamma) = k \mid R(\gamma) > 0) \tag{33}$$

$$= \text{Pr}(H = 0 \mid P \leq \gamma) \cdot 1 \tag{34}$$

$$= \text{Pr}(H = 0 \mid P \leq \gamma). \tag{35}$$

□

For large m it doesn't make much difference whether we regard the H_i as being random due to the Strong law of large numbers. Using Theorem 4 and Bayes' theorem we can derive

$$\text{pFDR}(\gamma) = \text{Pr}(H = 0 \mid P \leq \gamma) \tag{36}$$

$$= \frac{\text{Pr}(H = 0) \cdot \text{Pr}(P \leq \gamma \mid H = 0)}{\text{Pr}(P \leq \gamma)} \tag{37}$$

Using notation $\pi_0 = \Pr(H = 0)$ and the nesting property Eq. 3 it follows that P is uniformly distributed on $(0, 1)$ when H_{0i} is in fact true for each $i = 1, ..., m$ and we finally get:

$$\text{pFDR}(\gamma) = \frac{\pi_0 \cdot \gamma}{\Pr(P \leq \gamma)}. \tag{38}$$

To estimate pFDR (or FDR), we need estimators of π_0 and $\Pr(P \leq \gamma)$. Storey's estimator $\widehat{\pi}_0(\lambda)$ of π_0 is based on the following reasoning. As before let $p_1, ..., p_m$ be the observed p-values. Let $W(\lambda) = \#\{p_i > \lambda\}$ be the number of p-values greater than some value of λ. A large majority of p-values in the interval $[\lambda, 1]$, for λ not too small, should correspond to the true null hypotheses, and thus come from the uniform distribution on $[0, 1]$. This implies an expected value of $W(\lambda)$ to be approximately equal to the product $m \cdot \pi_0 = m_0$ and the length of the interval $[\lambda, 1]$:

$$\text{E}[W(\lambda)] \approx m \cdot \pi_0 \cdot (1 - \lambda) \tag{39}$$

and an estimator of π_0:

$$\widehat{\pi_0}(\lambda) = \frac{W(\lambda)}{m \cdot (1 - \lambda)} \tag{40}$$

Natural estimate of $\Pr(P \leq \gamma)$ is

$$\widehat{\Pr}(P \leq \gamma) = \frac{R(\gamma)}{m}. \tag{41}$$

Since pFDR is conditioned on $R(\gamma) > 0$ we need to divide our estimate of pFDR by $\Pr(R(\gamma) > 0)$. Lower bound for $\Pr(R(\gamma) > 0)$ is

$$1 - (1 - \gamma)^m. \tag{42}$$

Finally we get an estimate of pFDR as

$$\widehat{\text{pFDR}}_\lambda(\gamma) = \frac{W(\lambda) \cdot \gamma}{(1 - \lambda) \cdot \max\{R(\gamma), 1\} \cdot (1 - (1 - \gamma)^m)} \tag{43}$$

and since FDR is not conditioned on $R(\gamma) > 0$ we get an estimate of FDR as

$$\widehat{\text{FDR}}_\lambda(\gamma) = \frac{W(\lambda) \cdot \gamma}{(1 - \lambda) \cdot \max\{R(\gamma), 1\}}. \tag{44}$$

3 Power comparison of Benjamini-Hochberg procedure and Storey's direct approach

Let BH stand for Benjamini-Hochberg multiple-testing procedure and let ST stand for direct approach using Storey's estimates described in Section 2.

Proposition 5. *The detection power of BH procedure is always less or equal to the detection power of direct approach using Storey's estimates.*

Proof. Let $p_{(1)}, ..., p_{(m)}$ be observed ordered p-values and say that by using Storey's direct approach, we reject l hypotheses corresponding to the first l ordered observed p-values:

$$l = \max\{k : p_{(k)} \leq \lambda\}. \tag{45}$$

Using Storey's estimate of the proportion $\widehat{\pi}_0$ of true null hypotheses we estimate false discovery rate of direct approach: $\widehat{\text{FDR}}_{\text{ST}}$. We then use BH procedure with control at level $\widehat{\text{FDR}}_{\text{ST}}$ and estimate the rejection region:

$$\hat{k} = \max\{k : p_{(k)} \leq \frac{k}{m} \cdot \widehat{\text{FDR}}_{\text{ST}}\} \tag{46}$$

$$= \max\{k : p_{(k)} \leq \frac{k}{m} \cdot \frac{\widehat{\pi}_0 \cdot \gamma}{l/m}\} \tag{47}$$

$$= \max\{k : p_{(k)} \leq \frac{k \cdot \widehat{\pi}_0}{l} \cdot \gamma\}. \tag{48}$$

Since $\widehat{\pi}_0 \leq 1$ it follows that:

$$\hat{k} \leq l. \tag{49}$$

More precisely the proportion of hypotheses rejected by BH procedure of all the hypotheses rejected by Storey's direct approach is $\widehat{\pi}_0$. □

From this proof we can see that increase in detection power of Storey's direct approach over detection power of BH procedure depends on π_0 and at the same time on Storey's estimator $\widehat{\pi}_0$, i.e. even if π_0 is close to 0, there will be no improvement in detection power, if $\widehat{\pi_0}$ is close to 1.

We replicate the simulation study in [1] to compare detection power of BH procedure with detection power of direct approach using Storey's estimates. We fix rejection region and tuning parameter in the following way.

- Rejection region $\Gamma = [0, 0.01]$ or equivalently threshold $\gamma = 0.01$ for direct approach.
- The value of tuning parameter $\lambda = 0.5$ for Storey's estimation of $\widehat{\text{FDR}}_{ST}$ of false discovery rate FDR_{ST} of direct approach.

To put the two methods on equal ground for comparison, we use Storey's estimates to get an estimate $\widehat{\text{FDR}}_{ST}$ of the false discovery rate. Then use BH procedure to control the false discovery rate FDR_{BH} of BH procedure at the level $\widehat{\text{FDR}}_{\text{ST}}$.

We do $m = 1000$ one sided tests of null hypotheses where, for $i = 1, ..., m$, we test $H_{0i} : \mu = 0$ versus $H_{1i} : \mu > 0$ on the basis of p-values from simulated random variables $Z_i \sim N(\mu, 1)$, for $i = 1, ..., m$.

We limit the proportion π_0 of null hypotheses $Z_i \sim N(0, 1)$ to:

$$\pi_0 \in \{0.1, ..., 0.9\}. \tag{50}$$

The alternative distribution is $Z_i \sim N(2, 1)$.

The test statistics are in the form of observed p-values:

$$p_i = 1 - \Phi(z_i), \tag{51}$$

where z_i is the observed value of Z_i.

Let $m_0 = \pi_0 \cdot m$ be the number of true null hypotheses, $m_1 = m - m_0$ number of false null hypotheses and S the number of in fact false null hypotheses which are rejected. We estimate the *average detection power* of both methods:

$$\text{Power}_\phi = \frac{\text{E}(S)}{m - m_0}, \qquad \phi \in \{\text{ST}, \text{BH}\}. \tag{52}$$

In this setup, the probability of rejection $\Pr(R > 0)$ is very close to 1 for both methods. This means that error measures pFDR and FDR are very close to being equal. In our simulation we rejected at least one null hypotheses in each iteration $(R > 0)$, so estimate $\Pr(R > 0)$ was 1 and we only report FDR.

Table 2 summarizes 1000 simulations with values almost identical to Table in [1]. We see that Storey's estimates are very close to their true values. Estimated expected value of Storey's estimate:

$$\text{E}\left(\widehat{\text{FDR}}_{\text{ST}}\right) \tag{53}$$

π_0	FDR_{ST}	FDR_{BH}	Power_{ST}	Power_{BH}	$\text{E}(\widehat{\text{FDR}}_{\text{ST}})$	$\text{E}(\widehat{\pi_0})$	$\text{E}(\widehat{\gamma}_{\text{BH}})$
0.1	0.003	0.000	0.372	0.074	0.004	0.141	0.0000
0.2	0.007	0.002	0.373	0.122	0.008	0.237	0.0010
0.3	0.011	0.004	0.372	0.164	0.013	0.332	0.0010
0.4	0.017	0.007	0.371	0.203	0.019	0.427	0.0020
0.5	0.026	0.014	0.372	0.235	0.027	0.522	0.0030
0.6	0.039	0.024	0.371	0.265	0.040	0.618	0.0040
0.7	0.059	0.041	0.373	0.294	0.060	0.713	0.0050
0.8	0.096	0.077	0.372	0.320	0.099	0.809	0.0070
0.9	0.194	0.174	0.372	0.343	0.199	0.904	0.0080

Table 2: Average detection power comparison of BH procedure and direct approach.

of FDR for direct approach is almost always within 0.1% of actual FDR and always conservative. Estimated expected value of Storey's estimator of π_0:

$$\text{E}(\widehat{\pi_0}) \tag{54}$$

is always very close and conservative. Estimated expected value of the estimator $\widehat{\gamma}$ of rejection region threshold $\gamma = 0.01$ when using Benjamini-Hochberg procedure is very conservative:

$$\text{E}(\widehat{\gamma}_{\text{BH}}) << 0.01 \tag{55}$$

in all the cases of $\pi_0 = 0.1, ..., 0.9$. Actual FDR for BH procedure is a lot less than control $\widehat{\text{FDR}}_{\text{ST}}$. In Figure 1, we see that for π_0 close to 0, direct approach has much more detection power on average than BH procedure, which takes m instead of an estimate of m_0.

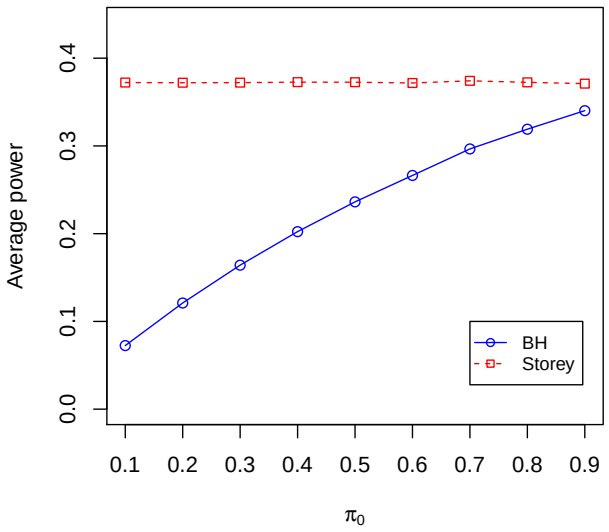

Figure 1: Average detection power comparison of BH procedure and direct approach.

4 Using Storey's estimator in Benjamini-Hochberg procedure

When using Storey's estimate $\widehat{\pi}_0$ of the proportion of true null hypotheses and $\widehat{\pi}_0 \cdot m$ instead of m in Benjamini-Hochberg procedure, the detection power of this improved Benjamini-Hochberg procedure is almost identical to Storey's direct approach.

Even more, we can show that it is always greater or equal to detection power of Storey's direct approach.

Proposition 6. *When using Storey's estimate $\widehat{\pi}_0$ in Benjamini-Hochberg procedure the detection power of Benjamini-Hochberg procedure is always greater or equal to detection power of Storey's direct approach.*

Proof. Let $p_1, ..., p_m$ be observed ordered p-values and let's say that by using direct approach, we reject l hypotheses corresponding to the first l ordered observed p-values, so:

$$l = \max\{k : p_{(k)} \leq \gamma\}. \tag{56}$$

We then use Storey's method to estimate proportion $\widehat{\pi}_0$ of true null hypotheses and from this estimate of false discovery rate: $\widehat{\text{FDR}}_{\text{ST}}$. We then use BH procedure with control at level $\widehat{\text{FDR}}_{\text{ST}}$ and $\widehat{\pi}_0 \cdot m$ instead of m and calculate:

$$\hat{k} = \max\{k : p_{(k)} \leq \frac{k}{\widehat{\pi}_0 \cdot m} \cdot \widehat{\text{FDR}}_{\text{ST}}\} \tag{57}$$

$$= \max\{k : p_{(k)} \leq \frac{k}{\widehat{\pi}_0 \cdot m} \cdot \frac{\widehat{\pi}_0 \cdot \gamma}{l/m}\} \tag{58}$$

$$= \max\{k : p_{(k)} \leq \frac{k \cdot \gamma}{l}\}. \tag{59}$$

and because $p_{(l)} \leq \gamma$ it follows that:

$$\hat{k} \geq l. \tag{60}$$

So the set of hypotheses rejected by direct approach are contained in the set of hypotheses rejected by this improved BH procedure. □

For this simulation we use the same setup as in the first but in BH procedure we use Storey's estimate $\widehat{\pi}_0$ of the proportion of true null hypotheses and $\widehat{\pi}_0 \cdot m$ instead of m. From summary of the results in Table 3 and from Figure 2 we see that estimated average detection power of BH procedure is always greater than estimated average power of direct approach using Storey's estimators:

$$\text{Power}_{\text{BH}} \geq \text{Power}_{\text{ST}}. \tag{61}$$

π_0	FDR_{ST}	FDR_{BH}	Power_{ST}	Power_{BH}	$\text{E}(\widehat{\text{FDR}}_{\text{ST}})$	$\text{E}(\widehat{\pi_0})$	$\text{E}(\widehat{\gamma}_{\text{BH}})$
0.1	0.003	0.003	0.372	0.373	0.004	0.141	0.0100
0.2	0.007	0.007	0.372	0.373	0.008	0.237	0.0100
0.3	0.011	0.011	0.372	0.373	0.013	0.332	0.0100
0.4	0.018	0.018	0.373	0.374	0.019	0.428	0.0100
0.5	0.026	0.026	0.373	0.374	0.027	0.523	0.0100
0.6	0.039	0.039	0.372	0.373	0.040	0.619	0.0100
0.7	0.057	0.058	0.374	0.376	0.060	0.713	0.0100
0.8	0.098	0.099	0.372	0.375	0.099	0.808	0.0100
0.9	0.193	0.197	0.371	0.377	0.200	0.906	0.0100

Table 3: Average detection power comparison of BH procedure using $\widehat{\pi}_0 \cdot m$ instead of m and direct approach.

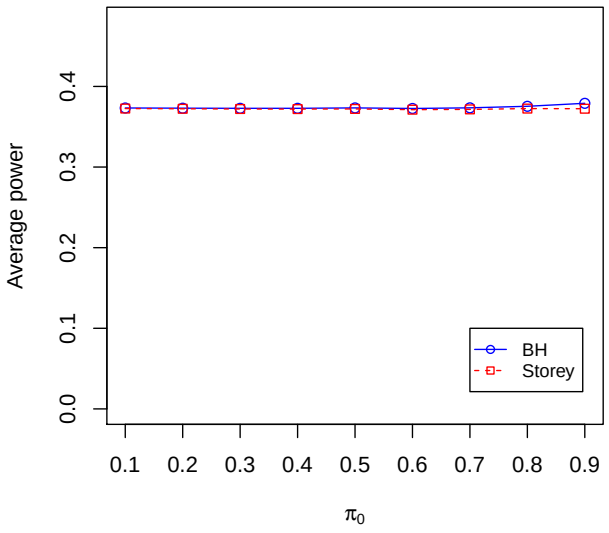

Figure 2: Average detection power comparison of BH procedure using $\widehat{\pi}_0 \cdot m$ instead of m and direct approach.

5 Properties of the Storey's estimator

In the previous two sections we showed that the estimator of π_0 plays a key role in detection power and in the estimation of error measure (p)FDR in the case of direct approach or equivalently in the level of control of FDR in the case of improved BH procedure when using estimator of π_0. Here we inspect some properties of this estimator.

Storey's estimator of the proportion of null hypotheses π_0 is:

$$\widehat{\pi_0}(\lambda) = \frac{W(\lambda)}{m \cdot (1-\lambda)}, \tag{62}$$

where $W(\lambda) = \#\{p_i > \lambda\}$ is the number of p-values greater than some value of parameter λ.

Proposition 7. *For any fixed value of λ Storey's estimator $\widehat{\pi}_0$ at most overestimates π_0 in expectation:*

$$E(\widehat{\pi}_0) \geq \pi_0. \tag{63}$$

Proof. Let $U(\lambda) = \#\{\text{null } p_i > \lambda\}$ be the number of p-values from in fact true null hypotheses which exceeds λ and $T(\lambda) = \#\{\text{alternative } p_j > \lambda\}$ corresponding number for the in fact false null hypotheses. Then $W(\lambda) = U(\lambda) + T(\lambda)$. If there are m_0 in fact true null hypotheses, then:

$$U \sim \text{Binomial}(m_0, 1-\lambda) \tag{64}$$

because $\Pr(P \geq \lambda \mid H_0) = 1 - \lambda$ and

$$T \sim \text{Binomial}\left(m - m_0, \Pr(P > \lambda \mid H_1)\right). \tag{65}$$

Putting all this together we get:

$$\mathrm{E}(\widehat{\pi}_0) = \frac{\mathrm{E}\left[U(\lambda)\right] + \mathrm{E}\left[T(\lambda)\right]}{m \cdot (1-\lambda)} \tag{66}$$

$$= \frac{m_0 \cdot (1-\lambda) + (m - m_0) \cdot \mathrm{Pr}(P > \lambda \mid H_1)}{m \cdot (1-\lambda)} \tag{67}$$

$$= \pi_0 + (1-\pi_0) \cdot \frac{\mathrm{Pr}(P > \lambda \mid H_1)}{1-\lambda} \tag{68}$$

$$\geq \pi_0. \tag{69}$$

□

This inequality is guaranteed only in expectation and largely depends on a good choice of λ, as can be seen from the plot in Figure 3, where we choose the actual proportion of true null hypotheses to be 0.9:

$$\pi_0 = 0.9 \tag{70}$$

and vary the value of λ in the range of $(0, 1)$.

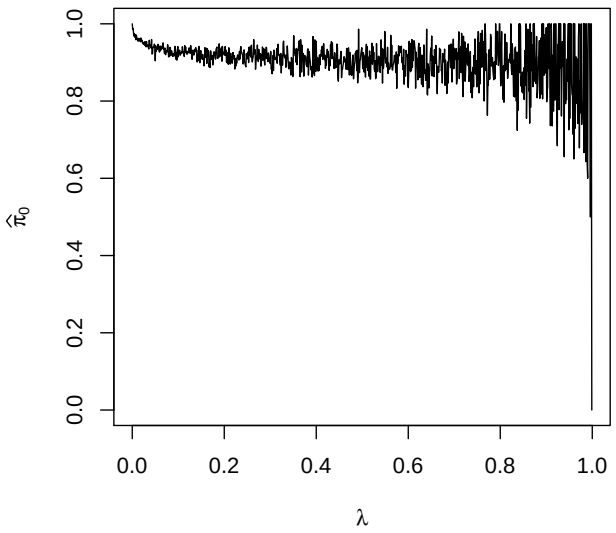

Figure 3: Storey's estimate $\widehat{\pi}_0$ of $\pi_0 = 0.9$ depending on the value of the parameter λ.

We introduce a new parameter μ_1 which is the distance between the distributions associated with H_0 and H_1. We use the same simulation setup described in Section 3 and simultaneously do $m = 1000$ one sided tests of null hypotheses, where the test statistics are in the form of observed p-values:

$$p_i = 1 - \Phi(z_i), \tag{71}$$

where z_i is the observed value of Z_i under null hypothesis:

$$Z_i \mid H_0 \sim N(0, 1) \tag{72}$$

and under alternative hypothesis:

$$Z_i \mid H_1 \sim N(\mu_1, 1). \tag{73}$$

The estimates of $\mathrm{E}(\widehat{\pi}_0)$ over the range of π_0 depending on the values of μ_1 and λ are summarized in Table 4. We see that the estimated upward bias increases, especially when π_0 is small, if we decrease the distance between the distributions μ_1 from 2 to 1. This is true in general and follows from Equation 68 where we see the upward bias increases when probability of observing p-values corresponding to alternative hypothesis H_1 that are greater than λ, increases. This estimated upward bias is reduced if we increase λ, in this case from 0.5 to 0.9. This does not obviously follow from Equation 68 and it is hard to say anything in general (maybe in the limit when $\lambda \to 1$).

This same upward bias is present in estimates of FDR and pFDR, since estimate $\widehat{\pi}_0$ appears in the numerator of Storey's estimates of FDR and pFDR. Because of this, we only report estimated $\mathrm{E}(\widehat{\pi}_0)$ from this simulation.

	$\mu_1 = 2$		$\mu_1 = 1$	
π_0	$\lambda = 0.5$	$\lambda = 0.9$	$\lambda = 0.5$	$\lambda = 0.9$
0.1	0.141	0.105	0.384	0.202
0.2	0.236	0.206	0.454	0.290
0.3	0.332	0.306	0.522	0.378
0.4	0.428	0.402	0.590	0.468
0.5	0.523	0.504	0.658	0.559
0.6	0.618	0.601	0.727	0.643
0.7	0.715	0.700	0.795	0.735
0.8	0.808	0.798	0.864	0.827
0.9	0.906	0.900	0.930	0.908

Table 4: Estimates $\mathrm{E}(\widehat{\pi}_0)$ depending on the values of parameters μ_1 and λ.

6 Storey's bootstrap method for choosing the tuning parameter

Storey proposed [1] a bootstrap procedure for choosing the optimal value λ^B of parameter λ which minimizes mean squared error of $\widehat{\pi}_0$. Using the same setup as before in Section 3 we simultaneously do 1000 simulation of 1000 one sided tests and in each simulation use 1000 bootstrap samples (samples of p-values with replacements) to estimate the optimal value λ^B. Results are in Table 5. We see the estimated expected value of this bootstrap estimator $\mathrm{E}(\widehat{\pi}_0^B)$ is very close to the actual π_0 for all π_0 in the range $0.1, .., 0.9$.

However in general the inequality:

$$\mathrm{E}(\widehat{\pi}_0) \geq \pi_0 \tag{74}$$

is no longer guaranteed for this estimator $\widehat{\pi}_0^B = \widehat{\pi}_0(\lambda^B)$. The value of λ is no longer fixed. In particular, choosing λ that is associated with the smallest $\widehat{\pi}_0$ (or with smallest $\widehat{\mathrm{pFDR}}$) will result in an underestimate of π_0 (and pFDR) as explained in [3]. Storey and Tibshirani have proposed a cubic spline approach to estimating π_0 to alleviate some of the difficulties associated with this bootstrap procedure.

π_0	$\mathrm{E}(\widehat{\pi}_0^B)$
0.1	0.108
0.2	0.207
0.3	0.308
0.4	0.404
0.5	0.505
0.6	0.603
0.7	0.703
0.8	0.804
0.9	0.901

Table 5: Estimates $\mathrm{E}(\widehat{\pi}_0^B)$.

7 Case of dependence

To simulate dependence, we introduce a new parameter ρ, which is the degree of dependence between the distributions associated with H_0 and H_1. We simultaneously do $m = 1000$ one sided test of null hypotheses, where we test $H_{0i} : \mu = 0$ versus $H_{1i} : \mu > 0$ on the basis of p-values from simulated dependent random variables Z'_i, for $i = 1, ..., m$ and:

$$Z'_i \sim Z_i + M, \tag{75}$$

where M is random variable independent of Z_i for each i and normally distributed with parameters:

$$M \sim N(0, \rho) \tag{76}$$

and Z_i is, as in Section 3, distributed:

$$Z_i \sim \begin{cases} N(0,1) & \text{if } H_{0i} \text{ is in fact true,} \\ N(2,1) & \text{if } H_{1i} \text{ is in fact true.} \end{cases}$$

and proportion of in fact true null hypotheses H_{0i} is π_0 out of all m hypotheses.

From this we can calculate degree of dependence in terms of ρ:

$$\mathrm{Cov}(Z_i + M, Z_j + M) = \begin{cases} \mathrm{Cov}(M, M) = \rho^2 & \text{if } i \neq j, \\ \mathrm{Var}(Z_i + M) = 1 + \rho^2 & \text{if } i = j \end{cases}$$

and

$$\mathrm{Cor}(Z_i + M, Z_j + M) = \begin{cases} \frac{\rho^2}{1+\rho^2} & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

On Figure 4 we plotted upward bias of Storey's estimator $\widehat{\mathrm{FDR}}_{\mathrm{ST}}$ if we fix $\pi_0 = 0.9$ and vary ρ in range from 0 to 0.9. And in Table 6 are summarized estimated Storey's estimators and estimated $\mathrm{E}(\widehat{\gamma}_{\mathrm{BH}})$ of BH procedure if we fix $\rho = 0.6$ and vary π_0. From Figure we see that degree of upward bias is very much influenced by the degree of dependence:

$$\mathrm{E}(\widehat{\mathrm{FDR}}_{\mathrm{ST}}) >> \mathrm{FDR}_{\mathrm{ST}}, \qquad \text{for } \rho > 0. \tag{77}$$

For example in our simulation for $\rho = 0.6$ or:

$$\mathrm{Cor}(Z'_i, Z'_j) = \frac{0.6^2}{1 + 0.6^2} \approx 0.26, \qquad i \neq j, . \tag{78}$$

and $\pi_0 = 0.9$, we have:

$$\mathrm{E}(\widehat{\mathrm{FDR}}_{\mathrm{ST}}) \approx 0.482 > 2 \cdot \mathrm{FDR}_{\mathrm{ST}} \approx 2 \cdot 0.228 = 0.456. \tag{79}$$

So our simulation is showing that in this case Storey's estimate of FDR is more than 2 times larger than actual FDR on average when using direct approach.

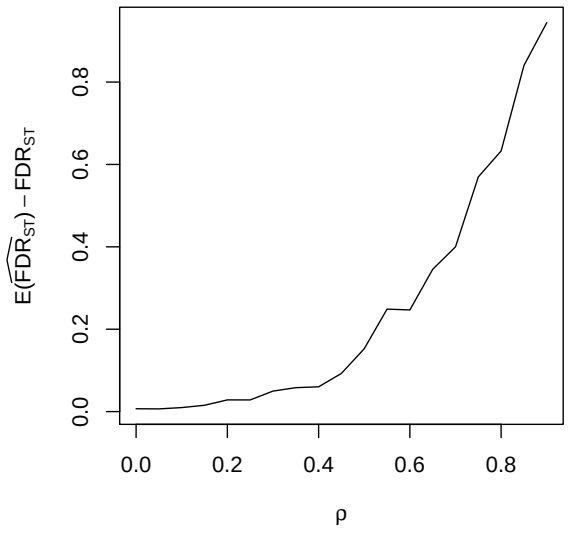

Figure 4: Upward bias of Storey's $\widehat{\text{FDR}}_{\text{ST}}$ depending on the value of the parameter ρ, where $\pi_0 = 0.9$.

π_0	FDR_{ST}	FDR_{BH}	Power_{ST}	Power_{BH}	$\text{E}(\widehat{\text{FDR}}_{\text{ST}})$	$\text{E}(\widehat{\pi_0})$	$\text{E}(\widehat{\gamma}_{\text{BH}})$
0.1	0.004	0.001	0.387	0.093	0.014	0.178	0.0010
0.2	0.010	0.002	0.386	0.142	0.027	0.275	0.0030
0.3	0.017	0.006	0.394	0.185	0.036	0.358	0.0040
0.4	0.024	0.010	0.380	0.213	0.049	0.464	0.0050
0.5	0.038	0.018	0.396	0.257	0.079	0.536	0.0090
0.6	0.055	0.030	0.398	0.287	0.082	0.621	0.0100
0.7	0.077	0.052	0.387	0.316	0.136	0.730	0.0180
0.8	0.128	0.103	0.394	0.351	0.242	0.810	0.0300
0.9	0.228	0.256	0.390	0.428	0.482	0.908	0.0910

Table 6: Upward bias of estimators in case of dependence $\rho = 0.6$.

8 Conclusions

We illustrate the duality between Storey's direct approach and Benjamini-Hochberg (BH) procedure when using Storey's estimate $\widehat{\pi}_0$ of proportion of null hypotheses π_0 in BH procedure. The performances of the two methods are almost identical in our simulations. This demonstrates that the estimator of π_0 plays a key role in multiple hypotheses testing. We show that Storey's estimator $\widehat{\pi}_0$ can be very upward biased when distance between distribution under null hypothesis H_0 and distribution under alternative hypotheses H_1 is small. In our simulations this upward bias of $\widehat{\pi}_0$ is reduced if we increase the value of λ and especially by tuning the parameter λ using Storey's bootstrap method. However this bootstrap method can result in underestimation of π_0. In case of dependence, we show by using simulation that dependence can lead to a very high overestimation of FDR.

References

[1] J. D. Storey, "A direct approach to false discovery rates," *Journal of the Royal Statistical Society: Series B (Statistical Methodology). Vol. 64(3)*, 2002.

[2] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society, Series B. 57(1)*, 1995.

[3] M. A. Black, "A note on the adaptive control of false discovery rates," *Journal of the Royal Statistical Society Series B (Statistical Methodology) 66(2):297-304*, 2004.

[4] J. D. Storey and R. Tibshirani, "Estimating the positive false discovery rate under dependence, with applications to dna microarrays," *Stanford Statistics Technical Report. No. 2001-28*, 2001.

[5] R. Heller, "False discovery rate control in multiple testing problems." `http://www.statistics.org.il/wp-content/uploads/2013/04/FDR-Ruth-Heller.pdf`, 2013. Accessed: 2019-10-24.

[6] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics, Springer, 2009.

[7] J. D. Storey, "The positive false discovery rate: A bayesian interpretation and the q-value," *The Annals of Statistics: Vol. 31(6)*, 2003.

[8] M. Langaas, B. H. Lindqvist, and E. Ferkingstad, "Estimating the proportion of true null hypotheses, with application to dna microarray data," *Journal of the Royal Statistical Society. Series B. Vol. 67(4)*, 2005.

[9] O. Oyeniran and H. Chen, "Estimating the proportion of true null hypotheses in multiple testing problems," *Journal of Probability and Statistics. Vol. 2016*, 2016.