This is a manuscript. It has not yet gone through peer review.

It is a revised version of:

Jonsson, Fredrik
URL: [http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-161740](http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-161740)
Applying the Benjamini–Hochberg procedure to a set of generalized $p$-values

Fredrik Jonsson, Uppsala University

Abstract

Consider $p$-values $\hat{p}_1, \ldots, \hat{p}_n$ corresponding to some “two-tailed” testing of one-dimensional point hypotheses, and the procedure determining the number of rejected hypotheses by $N = \max\{i : \hat{p}(i) \leq i\alpha/n\}$, given an ordering $\hat{p}(1) \leq \cdots \leq \hat{p}(n)$. Benjamini and Hochberg [1] showed, for the case of independent $p$-values, that the corresponding false discovery rate is bounded by $\alpha m_0/n$, with $m_0$ referring to the number of true null hypotheses. Later, Benjamini and Yekutieli [6, Corollary 3] extended this description regarding the control of Type III errors, showing for instance that the expected proportion of false directional conclusions is bounded by $\alpha/2$ in situations where none of the point hypotheses are true, and Type III errors are controlled at level $\alpha/2$ individually. With the ambition of capturing these and other related results within a unified and generalized framework, we consider a setting of “generalized $p$-values” in the following, referring to a set of decision functions and loss functions. Finally, we demonstrate in a simulation study that several related procedures for FDR-control can not be extended in a similar manner.

1 Introduction

The well-known Benjamini–Hochberg procedure [1], also known as Simes’ procedure [22], assumes $p$-values $\hat{p}_1, \ldots, \hat{p}_n$ and a significance level $\alpha$. The procedure then rejects hypotheses corresponding to the $N$ smallest values in an ordering $\hat{p}(1) \leq \cdots \leq \hat{p}(n)$, with $0 \leq N \leq n$ determined by $N = \max\{i : \hat{p}(i) \leq i\alpha/n\}$. Simes’ proved [22], under the assumption that $\hat{p}_1, \ldots, \hat{p}_n$ are independent random variables, that the probability of at least one rejection is bounded by $\alpha$ in the intersection of the null hypotheses.

Simes’ result was later extended by Benjamini and Hochberg [1], proving that the procedure generally — under the assumption of independence — controls the false discovery rate (the expected proportion of false rejections among the total number of rejections) at level $\alpha$. More precisely, it was shown that

$$FDR \leq \alpha m_0/n,$$

(1)

with $0 \leq m_0 \leq n$ referring to the number of true null hypotheses in a given situation.

It is well-known (cf. for instance the review [21] by Shaffer) that conventional significance testing of one-dimensional point hypotheses often may be conceived in a


Keywords and phrases. Multiple testing; false discovery rate; the Benjamini–Hochberg procedure; significance testing; directional errors; decision theory.

Date. November 22, 2011
slightly extended three-decision setting, with an appropriate directional conclusion if the hypothesis is rejected. This leads to additional “Type III” errors of making the reversed directional conclusion in situations where the hypothesis is false. For many test procedures, control of Type I errors at level $\alpha$ corresponds to control of Type III errors at level $\alpha/2$. It is also known that this correspondence is lost for most stepwise multiple procedures for control of familywise error rates (cf. e.g. [21] and [8]).

Perhaps somewhat surprisingly, Benjamini and Yekutieli [6, Corollary 3] showed that the given correspondence prevails for the Benjamini–Hochberg procedure with independent $p$-values. More precisely, it was shown that

$$FDR \leq \frac{\alpha}{n} \left( m_0 + \frac{1}{2}(n - m_0) \right),$$

if Type III errors are taken into account when testing one-dimensional point hypotheses.

There exists several, slightly different frameworks for three-decision procedures for directional conclusions, as one may see for instance from [21], [14] and [15]. With the ambition of capturing the results (1) and (2) — and other related formulations — within a unified and generalized formulation, we consider a setting of “generalized $p$-values” in the following.

Starting with a loss function $L$, corresponding to some decision function $\delta$, we assume that its expected loss can be controlled at level $\alpha$ through an indicator random variable $I$. In other words, we assume that

$$E_{\theta}(L(\delta, \theta) I) \leq \alpha,$$

for all $\theta \in \Theta$. A loss function $L = I\{\theta \in \Theta_0\}$ here corresponds to level-$\alpha$ testing of $\Theta_0$, with a confirmation $I = 1$ referred to as a rejection of $\Theta_0$. Following this path of generalization, we refer to a non-negative random variable $\hat{R}$ as a generalized $p$-value, with respect to $L$ and $\delta$, in case the following condition is fulfilled

$$E_{\theta}(L(\delta, \theta) I\{\hat{R} \leq \alpha\}) \leq \alpha,$$

for all $\alpha > 0$ and $\theta \in \Theta$.

A connection between these generalized $p$-values and the three-decision procedures for directional conclusions in [14] and [15] is given in Section 2.1.

In the setting of multiple loss functions $L_1, \ldots, L_n$ and corresponding confirmative indicators $I_1, \ldots, I_n$, note that the false discovery proportion (the proportion of false rejections among the total number of rejections), naturally generalizes into the following average loss among confirmed decisions,

$$\mathcal{L} = (L_1 I_1 + \cdots + L_n I_n) / (I_1 + \cdots + I_n),$$

(3)

(interpreted as zero if $I_1 = \cdots = I_n = 0$, that is, if no decisions are confirmed).

We show in Section 3 (with a more detailed description) that the Benjamini–Hochberg procedure applied to a set of generalized $p$-values $\hat{R}_1, \ldots, \hat{R}_n$ controls the expected average loss among confirmed decisions, in the sense that

$$E_{\theta}(\mathcal{L}) \leq \alpha,$$

assuming pairwise independence for $(\delta_1, \hat{R}_1), \ldots, (\delta_n, \hat{R}_n)$. Also, the Benjamini–Hochberg procedure is compared with more recent FDR-controlling procedures in a simulation study in Section 4, concluding that the FDR-control imposed by several “adaptive” procedures does not extend to the generalized setting.
1.1 General dependence

Several authors have considered the conservativeness of the Benjamini–Hochberg procedure with respect to control of FDR in cases of dependent \( p \)-values \( \hat{p}_1, \ldots, \hat{p}_n \).

For instance, Simes [22] included a simulation study concerning correlated test statistics and the intersection of the null hypotheses (where the notions of FDR and FWER coincide). Later, Samuel-Cahn [18] and Hochberg and Rom [12] gave examples with negatively dependent test statistics where conservativeness is violated, albeit only to a small extent, in the intersection of the null hypotheses. Earlier, Hommel [13] had proved (work which was extended by Falk [7]) that the probability of a false rejection is bounded by

\[
\alpha^* = \min\left\{1, \alpha \sum_{k=1}^{n} k^{-1}\right\}. 
\]

Moreover, it was claimed in [13] that examples demonstrating the sharpness of the bound could be given. Sarkar [19] proved that Simes’ original result on conservativeness for independent test statistics can be extended to a class of positively dependent test statistics.

Regarding the complement of the intersection of the null hypotheses, Benjamini and Yekutieli [5] generalized Hommel’s bound (4) in proving that

\[
\text{FDR} \leq \alpha \frac{m}{n} \sum_{k=1}^{n} k^{-1},
\]

with \( m \) referring to the number of true null hypotheses. Also, conservativeness was extended in [5] to a class of positively dependent test statistics (cf. also Sarkar [20] and Finner, Dickhaus and Roters [10, Section 4]). With an asymptotic perspective (regarding the number of null hypotheses), Finner, Dickhaus and Roters [9] investigated, theoretically and numerically, the impact of dependence on FDR for some commonly encountered test statistics.

In Section 3 we generalize the bounds of Hommel (4) and Benjamini–Yekutieli (5) by deriving corresponding bounds for \( E_{\theta}(L) \), increased by the factor \( \sum_{k=1}^{n} k^{-1} \) and valid without any assumption on dependence.

2 Risk-controlled decisions

Following the basic set-up in statistical decision theory, consider a random element \( X \) and a set of probability distributions \( \{P_{\theta} : \theta \in \Theta\} \), where each element is a possible law of \( X \) (cf. for instance [16, Section 1.1] or [17, Chapter 1]). Also, consider a non-negative loss function \( L \) with respect to some decision function \( \delta \) and the parameter space \( \Theta \) (cf. [17, Chapter 3]).

In this setting, assume that an indicator random variable \( I \) (a zero-one function of \( X \)) is given, with the purpose of confirming \( \delta \) at maximal risk \( \alpha \). Thus, \( \delta \) is confirmed whenever \( I = 1 \). Moreover, it is assumed that the following condition is fulfilled:

\[
E_{\theta}(L \cdot I) \leq \alpha, \quad \text{for all } \theta \in \Theta.
\]
Note that condition (6) generalizes the concept of a statistical test. Indeed, given some null hypothesis $\Theta_0 \subset \Theta$, and the following zero-one loss function,
\[ L = I\{\theta \in \Theta_0\}, \] (7)
condition (6) characterizes that the indicator $I$ serves as a level-$\alpha$ test of $\Theta_0$ (with $I = 1$ referred to as a rejection of $\Theta_0$). Thus, condition (6) generalizes statistical testing in two senses. First, losses are allowed to assume other values than 0 and 1. Second, losses are allowed to depend on some decision function $\delta$, i.e. on the random data $X$.

Next, generalizing the treatment of $p$-values in [16, Section 3.3], assume that some family $\{I(\alpha)\}_{0 \leq \alpha < \infty}$ of indicator random variables is given. Let the family be stochastically non–decreasing, in the sense that $I(\alpha) = 1$ implies $I(\alpha') = 1$ for $\alpha < \alpha'$, and assume that
\[ E_{\theta}(L \cdot I(\alpha)) \leq \alpha, \quad \text{for all } \theta \in \Theta \text{ and } 0 \leq \alpha < \infty. \] (8)
Then, by defining a random variable
\[ \hat{R} = \inf\{\alpha : I(\alpha) = 1\}, \]
condition (8) transfers into the following condition (cf. [16, Lemma 3.3.1]):
\[ E_{\theta}(LI(\hat{R} \leq \alpha)) \leq \alpha, \quad \text{for all } \theta \in \Theta \text{ and } 0 \leq \alpha < \infty. \] (9)
For brevity, we refer to $\hat{R}$ in (9) as a generalized $p$-value with respect to $L$ and $\delta$.

### 2.1 Directional conclusions

As an illustration of the generalized perspective expressed by condition (9), we follow the treatment of three-decision procedures for directional conclusions in [14] and [15]. Thus, assume that a partitioning $\{\Theta_1, \Theta_2\}$ of the parameter space $\Theta$ is given. For instance, $\Theta_1$ and $\Theta_2$ may refer to the two possible “directions” of a treatment effect. Then, being interested in controlled conclusions as to whether $\theta \in \Theta_1$ or $\theta \in \Theta_2$ holds, first consider naive decision functions $\delta$ into the binary decision space $\{1, 2\}$. A natural loss function for this decision, distinguishing between true and false conclusions, is then given by
\[ L(\delta, \theta) = I\{\theta \in \Theta_1\}I(\delta = 2) + I\{\theta \in \Theta_2\}I(\delta = 1). \] (10)
Introducing an indicator random variable $I$ fulfilling condition (6) in this context, we obtain a three-decision procedure by replacing $\delta$ with an indefinite decision in case $I = 0$. Moreover, the reversal rate is then controlled at level $\alpha$ (in the terminology of [14]).

Now, if $\hat{p}_1$ is a $p$-value for testing $\Theta_1$ and $\hat{p}_2$ is a $p$-value for testing $\Theta_2$, consider the directional decision function
\[ \delta = 2 \cdot I\{\hat{p}_1 \leq \hat{p}_2\} + 1 \cdot I\{\hat{p}_1 > \hat{p}_2\}. \]
In other words, $\delta$ concludes $\theta \in \Theta_1$ if it appears to be more profitable to reject $\Theta_2$, and vice versa. Moreover, consider
\[ \hat{R} = \min\{\hat{p}_1, \hat{p}_2\}. \]
Then, with $L$ given by (10),
\[
E_\theta(I \{ \hat{R} \leq \alpha \}) = E_\theta(I \{ \theta \in \Theta_1 \} I \{ \hat{p}_1 \leq \hat{p}_2 \} + I \{ \theta \in \Theta_2 \} I \{ \hat{p}_1 > \hat{p}_2 \}) = I \{ \theta \in \Theta_1 \} P_\theta(\hat{p}_1 \leq \alpha) + I \{ \theta \in \Theta_2 \} P_\theta(\hat{p}_2 \leq \alpha) \leq \alpha,
\]
which proves that $\hat{R}$ is an generalized $p$-value with respect to $L$ and $\delta$.

As a final remark, comparing the two loss functions in (7) and (10), note that both are zero-one loss functions. However, $L$ in (10) depends on given data.

2.2 Zero-one losses and the Bonferroni argument

Consider loss functions $L_1, \ldots, L_n$ with respect to one and the same statistical model $\{P_\theta : \theta \in \Theta\}$. Moreover, assume that corresponding generalized $p$-values $\hat{R}_1, \ldots, \hat{R}_n$ are given. Then, applying the indicator random variables $I_k = I \{ \hat{R}_k \leq \alpha \}$ simultaneously in order to confirm subsets of $\delta_1, \ldots, \delta_n$ may lead to some undesirable consequences.

For instance, assuming that $L_1, \ldots, L_n$ are zero-one loss functions (reporting the errors of $\delta_1, \ldots, \delta_n$ respectively), with $\theta$ given such that each pair $(L_k(\delta_k, \theta), \hat{R}_k)$ is independent of any other pair $(L_k'(\delta_k', \theta), \hat{R}_k')$, and $\alpha$ given such that the controlled risks satisfy
\[
E_\theta(L_1 I_1) = \cdots = E_\theta(L_n I_n) = \alpha,
\]
then, the probability of no error among confirmed decisions is given by,
\[
P_\theta\left(\bigcap_{k=1}^n \{ \min (L_k, I_k) = 0 \}\right) = \prod_{k=1}^n \left(1 - E_\theta(L_k I_k)\right) = (1 - \alpha)^n,
\]
which is considerably smaller than $1 - \alpha$ if $n$ is large.

As an alternative, consider replacing $I_k$ by the more restrictive “Bonferroni-indicators” $I'_k = I \{ \hat{R}_k \leq \alpha/n \}$. Then, assuming zero-one loss functions $L_1, \ldots, L_n$, the probabilities of at least one error among confirmed decisions satisfy the following bound:
\[
P_\theta\left(\bigcup_{k=1}^n \{ L_k I'_k = 1 \}\right) \leq \sum_{k=1}^n E_\theta(L_k I'_k) \leq n \cdot \alpha/n = \alpha.
\]

2.3 Maximal and average loss among confirmed decisions

Note that condition (11) refers to the fact that the expected maximal loss among confirmed decisions is controlled at level $\alpha$. In other words, $E_\theta(L_{max}) \leq \alpha$ for all $\theta \in \Theta$, with respect to
\[
L_{max} = \max \{ L_1 I_1, \ldots, L_n I_n \}.
\]
Here, $L_{max}$ might be considered as a combined measure of loss among confirmed decisions. An alternative measure is given by the following average loss among confirmed decisions,
\[
L = (L_1 I_1 + \cdots + L_n I_n)/(I_1 + \cdots + I_n),
\]
with \( \mathcal{L} \) interpreted as zero if no decisions are confirmed (\( I_1 = \cdots = I_n = 0 \)).

In multiple testing, obtaining \( \mathbb{E}_\theta(\mathcal{L}_{\text{max}}) \leq \alpha \) is referred to as controlling the FWER (family-wise error rate, cf. [16, p. 349]) at level \( \alpha \). Similarly, obtaining \( \mathbb{E}_\theta(\mathcal{L}) \leq \alpha \) is referred to as controlling the FDR (false discovery rate, cf. [1]) at level \( \alpha \).

### 2.4 The Benjamini–Hochberg procedure

Consider non-random, zero-one loss functions \( L_1, \ldots, L_n \) and corresponding \( p \)-values \( R_1, \ldots, R_n \), with respect to some statistical model \( \{ P_\theta : \theta \in \Theta \} \). Let \( R(1), \ldots, R(p) \) refer to \( R_1, \ldots, R_n \) in a given ordering \( R(1) \leq \cdots \leq R(p) \). Benjamini and Hochberg proved [1], assuming \( R_1, \ldots, R_n \) to be independent random variables and a level \( \alpha \), that the following indicator random variables (with \( \hat{R}(n+1) := \infty \) for notational convenience),

\[
I_k = \sum_{i=1}^{n} I\{n\hat{R}_k \leq \alpha i\} \prod_{j=i+1}^{n+1} I\{n\hat{R}_{(j)} > \alpha j\} I\{n\hat{R}_{(i)} \leq \alpha i\}, \quad k = 1, \ldots, n, \tag{14}
\]

controls the expected average loss among confirmed decisions at level \( \alpha \).

Note that Definition (14) is usually interpreted in a stepwise manner. Thus, starting at the bottom of the hierarchy of ordered \( p \)-values, \( \hat{R}(n) \) is first compared with \( \alpha \). If smaller than \( \alpha \), all decisions are confirmed. Otherwise, \( n\hat{R}(n-1) \) is compared with \( \alpha(n-1) \), confirming all decisions corresponding to \( \hat{R}(3), \ldots, \hat{R}(n-1) \) in case \( \alpha(n-1) \) exceeds \( n\hat{R}(n-1) \). Otherwise, \( n\hat{R}(n-2) \) is compared with \( (n-2)\alpha \), etcetera. Thus, the procedure confirms a random number \( N \), \( 0 \leq N \leq n \), of the given decisions, corresponding to the \( N \) smallest among \( \hat{R}_1, \ldots, \hat{R}_n \), with \( N \) determined by

\[
N = \max\{i : \hat{R}_{(i)} \leq i\alpha/n\},
\]

(setting \( \hat{R}_{(0)} := 0 \) for notational convenience).

Applying the definition (13) of \( \mathcal{L} \) as the average loss among confirmed decisions to the indicators (14), one obtains

\[
\mathcal{L} = \sum_{i=1}^{n} \frac{1}{i} I\{N = i\} \sum_{k=1}^{n} L_k I\{n\hat{R}_k \leq \alpha i\}
= \sum_{k=1}^{n} \frac{1}{i} L_k I\{n\hat{R}_k \leq \alpha i\} \prod_{j=i+1}^{n+1} I\{n\hat{R}_{(j)} > \alpha j\} I\{n\hat{R}_{(i)} \leq \alpha i\}.
\]

The fact that \( \mathbb{E}_\theta(\mathcal{L}) \leq \alpha \) can thus be expressed as

\[
\sum_{k=1}^{n} \sum_{i=1}^{n} \frac{1}{i} \mathbb{E}_\theta(L_k I\{n\hat{R}_k \leq \alpha i\} \prod_{j=i+1}^{n+1} I\{n\hat{R}_{(j)} > \alpha j\} I\{n\hat{R}_{(i)} \leq \alpha i\}) \leq \alpha. \tag{15}
\]

### 3 Main results

Theorem 3.1 generalizes the validity of the bound (15) to the setting of non-negative loss functions \( L_1, \ldots, L_n \) and corresponding generalized \( p \)-values \( \hat{R}_1, \ldots, \hat{R}_n \).
Part (i) generalizes the original result [1] and the treatment of directional errors in [6, Corollary 3]. In Part (i), the pairs \((\delta_1, \hat{R}_1), \ldots, (\delta_n, \hat{R}_n)\) are assumed to be independent. However, nothing is assumed about the dependence between \(\delta_k\) and \(\hat{R}_k\) within each given pair \((\delta_k, \hat{R}_k)\).

Note that, with ordinary testing of hypotheses and \(\theta\) given such that \(0 \leq m_0 \leq n\) of the \(n\) hypotheses are true, then, referring to condition (16),

\[
\sum_{k=1}^{n} \lambda_k = m_0.
\]

Also, if Type III errors are taken into account (cf. the discussion in Section 1 and [6, Corollary 3]), assumed to be controlled at level \(\alpha/2\) by the corresponding level-\(\alpha\) tests, then

\[
\sum_{k=1}^{n} \lambda_k = m_0 + \frac{1}{2}(n - m_0).
\]

Thus, if none of the point hypotheses is believed to be true, but Type III errors are taken into account, then

\[
\sum_{k=1}^{n} \lambda_k = n/2.
\]

Alternatively, one may adopt the closely related formulation in Section 2.1 for the ambition of making controlled directional conclusions, avoiding any reference to Type I errors (which usually amounts to scaling the usual \(p\)-values by a factor 1/2), leading to a general bound with

\[
\sum_{k=1}^{n} \lambda_k = n.
\]

Part (ii) generalizes Theorem 1.3 in [5]. In this case, a corresponding bound is given, increased by a factor

\[
\sum_{i=1}^{n} \frac{1}{i} \approx 0.58 + \log n,
\]

compared to the bound in Part (i), valid without any assumption on dependence.

**Theorem 3.1.** Let \(\delta_1, \ldots, \delta_n\) be decision functions, with associated loss functions \(L_1, \ldots, L_n\) on some statistical model \(\{P_\theta : \theta \in \Theta\}\). Moreover, let \(\theta \in \Theta\), non-negative random variables \(R_1, \ldots, R_n\), and non-negative numbers \(\lambda_1, \ldots, \lambda_n\) be given such that, for all \(\beta \leq \alpha\),

\[
E_{\theta}(L_k I\{\hat{R}_k \leq \beta\}) \leq \lambda_k \beta.
\]  

(16)

Consider the corresponding level-\(\alpha\) Benjamini–Hochberg procedure and let \(L\) refer to the average loss among confirmed decisions.
(i) Assume that the pairs \((\delta_1, \hat{R}_1), \ldots, (\delta_n, \hat{R}_n)\) are independent. Then,

\[
E_{\theta}(L) \leq \frac{1}{n} \sum_{k=1}^{n} \lambda_k.
\]  

Moreover, if equality holds in (16), for all \(k = 1, \ldots, n\) and all \(\beta \leq \alpha\), then equality also holds in (17).

(ii) Without any additional assumptions,

\[
E_{\theta}(L) \leq \left(\frac{1}{n} \sum_{k=1}^{n} \lambda_k\right) \left(\sum_{i=1}^{n} \frac{1}{i}\right). 
\]

Proof. To begin with, recall from (14)-(15) that \(E_{\theta}(L)\) is given by (with the convention \(\hat{R}_{(n+1)} = \infty\)):

\[
E_{\theta}(L) = \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{1}{i} E_{\theta}(L_k I\{n\hat{R}_k \leq \alpha i\} \prod_{j=i+1}^{n+1} I\{n\hat{R}_{(j)} > \alpha j\} I\{n\hat{R}_{(i)} \leq \alpha i\}).
\]  

For each integer \(k = 1, \ldots, n\), let \(X_{2,k} \leq \cdots \leq X_{n,k}\) denote the ordered random variables \(\hat{R}_1, \ldots, \hat{R}_n\) when \(\hat{R}_k\) is removed. Moreover, set \(X_{1,k} = 0\) and \(X_{n+1,k} = \infty\), for convenience. Thus, rewriting (18) we obtain

\[
E_{\theta}(L) = \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{1}{i} E_{\theta}(L_k I\{n\hat{R}_k \leq \alpha i\} \prod_{j=i+1}^{n+1} I\{nX_{j,k} > \alpha j\} I\{nX_{i,k} \leq \alpha i\}). 
\]  

Regarding part (i), note that \(X_{2,k} \leq \cdots \leq X_{n,k}\) are independent of the two random variables \(L_k(\delta_k, \theta)\) and \(\hat{R}_k\). Hence, (19) simplifies to

\[
E_{\theta}(L) = \sum_{k=1}^{n} \sum_{i=1}^{n} \frac{1}{i} E_{\theta}(L_k I\{n\hat{R}_k \leq \alpha i\}) E\left(\prod_{j=i+1}^{n+1} I\{nX_{j,k} > \alpha j\} I\{nX_{i,k} \leq \alpha i\}\right).
\]

Next, applying assumption (16) gives

\[
E_{\theta}(L) \leq \frac{1}{n} \sum_{k=1}^{n} \lambda_k \sum_{i=1}^{n} E\left(\prod_{j=i+1}^{n+1} I\{nX_{j,k} > \alpha j\} I\{nX_{i,k} \leq \alpha i\}\right),
\]  

with equality in case of equalities in (16). Now, for any \(k = 1, \ldots, n\),

\[
\sum_{i=1}^{n} \prod_{j=i+1}^{n+1} I\{nX_{j,k} > \alpha j\} I\{nX_{i,k} \leq \alpha i\} = 1,
\]

since \(0 = X_{1,k} \leq X_{2,k} \leq \cdots \leq X_{n+1,k} = \infty\). Thus, (20) simplifies to

\[
E_{\theta}(L) \leq \frac{1}{n} \sum_{k=1}^{n} \lambda_k,
\]
which completes the proof of part (i).

Regarding part (ii), we note in view of (21) that, for any \( k = 1, \ldots, n \),

\[
\sum_{i=1}^{n} \frac{1}{i} I\{nR_k \leq \alpha i\} \prod_{j=i+1}^{n+1} I\{nX_{j,k} > \alpha j\} I\{nX_{i,k} \leq \alpha i\} \\
\leq \max_{1 \leq i \leq n} \left\{ \frac{1}{i} I\{nR_k \leq \alpha i\} \right\} \sum_{i=1}^{n} \prod_{j=i+1}^{n+1} I\{nX_{j,k} > \alpha j\} I\{nX_{i,k} \leq \alpha i\} \\
= \max_{1 \leq i \leq n} \left\{ \frac{1}{i} I\{nR_k \leq \alpha i\} \right\}.
\]

(22)

We also note that

\[
\max_{1 \leq i \leq n} \left\{ \frac{1}{i} I\{nR_k \leq \alpha i\} \right\} = \sum_{i=1}^{n} \frac{1}{i} I\{\alpha (i-1) < nR_k \leq \alpha i\} \\
= \frac{1}{n} I\{\hat{R}_k \leq \alpha\} + \sum_{i=1}^{n-1} \left( \frac{1}{i} - \frac{1}{i+1} \right) I\{nR_k \leq \alpha i\} \\
= \frac{1}{n} I\{\hat{R}_k \leq \alpha\} + \sum_{i=1}^{n-1} \frac{1}{i(i+1)} I\{nR_k \leq \alpha i\}.
\]

(23)

Applying (22)-(23) in (19), it follows that

\[
E_\theta(L) \leq \sum_{k=1}^{n} E_\theta \left( L_k \left( \frac{1}{n} I\{\hat{R}_k \leq \alpha\} + \sum_{i=1}^{n-1} \frac{1}{i(i+1)} I\{nR_k \leq \alpha i\} \right) \right).
\]

(24)

Finally, applying assumption (16) to (24), we deduce that

\[
E_\theta(L) \leq \alpha \frac{1}{n} \sum_{k=1}^{n} \lambda_k \left( 1 + \sum_{i=1}^{n-1} \frac{1}{i+1} \right) = \left( \alpha \frac{1}{n} \sum_{k=1}^{n} \lambda_k \right) \left( \sum_{i=1}^{n} \frac{1}{i} \right),
\]

proving part (ii) of the theorem.

\[\square\]

4 A simulation study

As a simple statistical model for theoretical evaluations and comparisons of multiple testing procedures, consider Dirac-uniform configurations (referred to as “U1” in Table 1). In other words, let \( 0 \leq m \leq n \) be given, with \( m \) determining the number of true null hypotheses among the total number \( n \). Then, let \( \hat{p}_1, \ldots, \hat{p}_m \) refer to i.i.d. random variables uniformly distributed on the interval \([0, 1]\) and assume that the remaining \( p \)-values are constantly equal to zero.

We also consider a related model here referred to as “U1\( \frac{1}{2} \)”. In this case, with \( 0 \leq m \leq n \) given, let \( \hat{R}_1, \ldots, \hat{R}_m \) be i.i.d. random variables uniformly distributed on the interval \([0, 1/2]\), with associated decisions \( \delta_1, \ldots, \delta_m \) i.i.d. uniformly distributed on the binary set \( \{0, 1\} \). Let all decisions be evaluated with respect to the loss function \( L(\delta) = I\{\delta = 1\} \). The remaining generalized \( p \)-values \( \hat{R}_{m+1}, \ldots, \hat{R}_n \) and decision functions \( \delta_{m+1}, \ldots, \delta_n \) are assumed to be constantly equal to zero.
As a motivation of the $U^\frac{1}{2}$-model in this situation, consider the treatment of directional conclusions in Section 2.1. A simple example is then given by the family $\{N(\mu, 1) : \mu \in \mathbb{R}\}$ of unit variance, univariate normal distributions, divided into $H_1 : \mu \leq 0$ and $H_2 : \mu > 0$. Let the naive directional decision function be given by

$$\delta = 2 \cdot I\{X > 0\} + 1 \cdot I\{X \leq 0\}.$$  

Then, $\hat{R} = \Phi(-|X|)$ serves as a generalized $p$-value with respect to $\delta$ and the directional loss function,

$$L = I\{\mu \leq 0\}I\{X > 0\} + I\{\mu > 0\}I\{X \leq 0\}.$$  

Note that $\hat{R}$ is uniformly distributed on $[0, 1/2]$, and that $L(\delta)$ is uniformly distributed on $\{0, 1\}$ in case $\mu = 0$. Moreover, $\hat{R}$ and $\delta$ are then also independent. Thus, $U^\frac{1}{2}$ corresponds to directional conclusions with $\mu_1 = \cdots = \mu_m = 0$, $\mu_{m+1} = \cdots = \mu_n = \infty$, and independent observations.

Table 1 reports on FDR (expected proportion of false conclusions) regarding four multiple testing procedures applied to $\hat{R}_1, \ldots, \hat{R}_n$ in the above models $U_1$ and $U^\frac{1}{2}$. All procedures are performed at level $\alpha = 0.05$, with respect to $n = 100$ and $m = 2, 5, 20, 99$. Moreover, all procedures are known to control the FDR at level $\alpha = 0.05$ for $U_1$.

<table>
<thead>
<tr>
<th></th>
<th>$m = 2$</th>
<th>$m = 5$</th>
<th>$m = 20$</th>
<th>$m = 99$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$U_1$</td>
<td>$U^\frac{1}{2}$</td>
<td>$U_1$</td>
<td>$U^\frac{1}{2}$</td>
</tr>
<tr>
<td>BH</td>
<td>0.001</td>
<td>0.001</td>
<td>0.0025</td>
<td>0.0025</td>
</tr>
<tr>
<td>BL</td>
<td>0.020</td>
<td>0.010</td>
<td>0.050</td>
<td>0.025</td>
</tr>
<tr>
<td>GBS</td>
<td>0.016</td>
<td>0.010</td>
<td>0.034</td>
<td>0.025</td>
</tr>
<tr>
<td>$S^\frac{1}{2}$</td>
<td>0.010</td>
<td>0.010</td>
<td>0.025</td>
<td>0.025</td>
</tr>
</tbody>
</table>

Table 1: FDR for the two models $U_1$ and $U^\frac{1}{2}$ for $n = 100$ and selected values of $m$. BH refers to the Benjamini–Hochberg procedure, BL to the Benjamini–Liu procedure [4], GBS to the Gavrilov–Benjamini–Sarkar procedure [11], and $S^\frac{1}{2}$ to the Storey procedure [24] with $\lambda = 1/2$.

Note to begin with that the error rate of the Benjamini–Hochberg procedure is constant when comparing the two models $U_1$ and $U^\frac{1}{2}$ with $m$ fixed. Indeed, the expected proportion of false conclusions is given by $\alpha m/n$, as explained by Theorem 3.1 (i) above.

The step-down FDR-controlling procedure of Benjamini–Liu [4] is known to be more powerful compared to the Benjamini-Hochberg procedure when $m$ is small (cf. [4]). This is in agreement with Table 1, where it can be seen that BL produces larger false discovery rates for $m = 2$ and 5 compared to BH. Also, it appears from Table 1 that BL is conservative with respect to the model $U^\frac{1}{2}$. One may verify that the false discovery rates of BL applied to $U^\frac{1}{2}$ is approximately bounded by 0.025, for any $0 \leq m \leq 100$ in this case.

Finally, the two “adaptive” FDR-procedures which we here consider (cf. [10], [11] for GBS, and [23], [24] for $S^\frac{1}{2}$) sometimes produce false discovery rates much larger than 5% when applied to the $U^\frac{1}{2}$-model. This might be characteristic for adaptive FDR-controlling procedures (cf. also suggestions in [3] and [2]). Indeed, the notion of
adaptiveness refers to the strategy of incorporating an estimated value of the number of true null hypotheses into an existing step-wise procedure (cf. e.g. [3, Section 3]). This strategy may thus be closely related to the assumption of data-invariant, zero-one loss functions $L_1, \ldots, L_n$.

Comparing the procedures GBS and $S^*_1$, it was suggested in the final discussion of [11] that GBS might be less powerful compared to $S^*_1$ for control of Type I errors with independent test statistics, whereas $S^*_1$ might have a worse behaviour regarding control of FDR in cases of dependent test statistics. In Table 1, $S^*_1$ has a worse behaviour regarding the model $U^*_1$, notably for $m = 99$.

References


