Biometrika (2024), **111**, 4, *pp.* 1129–1150 https://doi.org/10.1093/biomet/asae018
Advance Access publication 23 March 2024

Flexible control of the median of the false discovery proportion

BY JESSE HEMERIK

Econometric Institute, Erasmus University,
Burg. Oudlaan 50, 3062 PA Rotterdam, The Netherlands
hemerik@ese.eur.nl

ALDO SOLARI

Department of Economics, Ca' Foscari University, Cannaregio 873, 30121 Venice, Italy

AND JELLE J. GOEMAN

Department of Biomedical Data Sciences, Leiden University Medical Center,
Einthovenweg 20, 2333 ZC Leiden, The Netherlands

SUMMARY

We introduce a multiple testing procedure that controls the median of the proportion of false discoveries in a flexible way. The procedure requires only a vector of p-values as input and is comparable to the Benjamini–Hochberg method, which controls the mean of the proportion of false discoveries. Our method allows free choice of one or several values of α after seeing the data, unlike the Benjamini–Hochberg procedure, which can be very anti-conservative when α is chosen post hoc. We prove these claims and illustrate them with simulations. The proposed procedure is inspired by a popular estimator of the total number of true hypotheses. We adapt this estimator to provide simultaneously median unbiased estimators of the proportion of false discoveries, valid for finite samples. This simultaneity allows for the claimed flexibility. Our approach does not assume independence. The time complexity of our method is linear in the number of hypotheses, after sorting the p-values.

Some key words: Control; Estimation; False discovery proportion; False discovery rate; Post hoc.

1. INTRODUCTION

Multiple hypothesis testing procedures have the common aim of ensuring that the number of incorrect rejections, i.e., false positives, is likely to be small. The most commonly used multiple testing procedures control either the familywise error rate or the false discovery rate, FDR (Dickhaus, 2014; Harvey et al., 2020). The false discovery rate is the expected value of the false discovery proportion, FDP, which is the proportion of false positives among all rejections of null hypotheses. Controlling the FDR means ensuring that the expected FDP is kept below some prespecified value α (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Goeman & Solari, 2014).

The FDP, which is an unknown quantity, can vary widely about its mean when the tested variables are strongly correlated (Efron, 2007; Schwartzman & Lin, 2011; Delattre & Roquain, 2015). For this reason, methods have been developed that do not control the FDR or estimate the FDP, but instead provide a confidence interval for the FDP (Hemerik & Goeman, 2018). Some methods provide confidence intervals for several choices of the set of rejected hypotheses that are simultaneously valid with high confidence (Genovese & Wasserman, 2004, 2006; Meinshausen, 2006; Hemerik et al., 2019; Blanchard et al., 2020; Katsevich & Ramdas, 2020; Goeman et al., 2021; Blain et al., 2022; Vesely et al., 2023). There are also procedures, including the methods just mentioned, that ensure that the FDP remains small with high confidence (van der Laan et al., 2004; Lehmann & Romano, 2005; Guo & Romano, 2007; Romano & Wolf, 2007; Farcomeni, 2008; Roquain, 2011; Guo et al., 2014; Delattre & Roquain, 2015; Ditzhaus & Janssen, 2019; Döhler & Roquain, 2020; Basu et al., 2023; Miecznikowski & Wang, 2023). This is often termed false discovery exceedance control.

Methods that ensure that the FDP remains small with high confidence can provide very clear and useful error guarantees. The downside of these methods, however, is that under dependence they often do not have sufficient power to reject any hypotheses, even if there is a substantial amount of signal in the data. The reason is that these methods require not merely that the FDP be small on average, but also that it be small with high confidence. As a result, users may prefer approaches with weaker guarantees, such as FDR methods. An alternative is to take $\alpha = 0.5$ in FDP methods.

The most popular FDR method is the Benjamini–Hochberg method (Benjamini & Hochberg, 1995). FDR methods generally require the user to choose α before looking at the data. Common choices for α are 0.05 and 0.1. The methods guarantee that the FDR is kept below α. However, researchers would often like to change α post hoc. For example, if no hypotheses are rejected for $\alpha = 0.05$, a researcher may want to increase α to 0.1, changing the FDP target in order to obtain more rejections. In other cases, the user will want to decrease α. However, as we show in § 3.2 and the Supplementary Material, choosing α post hoc can severely invalidate methods such as the Benjamini–Hochberg procedure. Moreover, the user may want to report results for several values of α, while providing a simultaneous error guarantee. There is a need for methods that allow these types of inference.

In this article, we introduce a class of multiple testing methods that allow us to choose the threshold freely after looking at the data. Our methodology requires only a vector of p-values as input and is nonasymptotic. Our procedure controls the median of the FDP rather than the mean. For this and other reasons, we denote the target FDP by $\gamma \in [0, 1]$ instead of α; this is inspired by Romano & Wolf (2007), Harvey et al. (2020) and Basu et al. (2023). Controlling the median means that the FDP is at most γ with probability at least 0.5. We will refer to this as mFDP control. We remark that mFDP control can also be achieved with several existing FDP methods, by taking $\alpha = 0.5$. Like some existing methods, our procedure is flexible in the sense that γ can be freely chosen after seeing the data. Further, the procedure is adaptive, in the sense that it does not necessarily become conservative if the fraction of false hypotheses is large. We prove that our procedure is valid under a novel type of assumption on the joint distribution of the p-values. In particular, our method does not require independence. Moreover, the method was found to be valid in all simulation settings considered. Further, we prove that the proposed procedures are often admissible, i.e., they cannot be uniformly improved upon (Goeman et al., 2021). Since the method of Goeman et al. (2019) is also flexible and admissible in some settings, we compare our method with that one in simulations. We also compare with the elegant and fast method of Katsevich & Ramdas (2020).

The proposed methodology has been implemented in the R (R Development Core Team, 2024) package mFDP, available on CRAN.

Our procedure is partly inspired by an existing estimator of the fraction $\pi_0 \in [0,1]$ of true hypotheses among all hypotheses. This estimator is mentioned in Schweder & Spjøtvoll (1982) and advocated in Storey (2002). We refer to it as the Schweder–Spjøtvoll–Storey estimator. Some publications refer to it as Storey's estimator or the Schweder–Spjøtvoll estimator (Hoang & Dickhaus, 2022). The literature contains multiple π_0 estimators based on p-values (Rogan & Gladen, 1978; Hochberg & Benjamini, 1990; Langaas et al., 2005; Meinshausen et al., 2006; Rosenblatt, 2021). As a side result of our investigation of π_0 and FDP estimation, we add to this literature a novel π_0 estimator that is slightly different from the Schweder–Spjøtvoll–Storey estimator, unless its tuning parameter is 0.5.

The proposed methodology also draws from an idea in Hemerik et al. (2019), which is to construct simultaneous FDP bounds, called confidence envelopes, in a manner that is partly data-based and partly reliant on a prespecified family of candidate envelopes. The simultaneity of the constructed bounds allows for post hoc selection of rejection thresholds and hence post hoc specification of γ. The methodology proposed here is applicable in many situations, where the method of Hemerik et al. (2019) is not. The reason is that one cannot generally use permutations if one only has p-values, which is the setting we assume.

Our mFDP-controlling approach conceptually relates to recent methods that bound the FDR by α by finding the largest p-value threshold for which some conservative estimate of the FDP is below α (Barber & Candès, 2015; Li & Barber, 2017; Lei & Fithian, 2018; Lei et al., 2021; Luo et al., 2022; Rajchert & Keich, 2022). Those methods do not offer the simultaneity provided in the present work.

2. Median unbiased estimation of the false discovery proportion

2.1. *Notation*

Throughout this paper we consider hypotheses $H_1, \ldots, H_m$ and corresponding p-values $p_1, \ldots, p_m$, which take values in $(0,1]$. Write $p = (p_1, \ldots, p_m)$. Let $\mathcal{N} = \{1 \leqslant i \leqslant m : H_i \text{ is true}\}$ be the set of indices of true hypotheses and let $N = |\mathcal{N}|$ be the number of true hypotheses, which we assume to be strictly positive for convenience. The fraction of true hypotheses is $\pi_0 = N/m$. Let $q_1, \ldots, q_N$ denote the p-values corresponding to the true hypotheses, in any order. Write $q = (q_1, \ldots, q_N)$.

If $t \in (0,1)$, we write $\mathcal{R}(t) = \{1 \leqslant i \leqslant m : p_i \leqslant t\}$. We call $\mathcal{R} = \mathcal{R}(t)$ the set of rejected hypotheses, since t will usually denote the p-value threshold. Write $R = |\mathcal{R}|$. Let $V = |\mathcal{N} \cap \mathcal{R}|$ be the number of true hypotheses in $\mathcal{R}$, i.e., the number of false positive findings. We write $a \wedge b$ for the minimum of the numbers a and b.

2.2. *The Schweder–Spjøtvoll–Storey estimate*

Our first results, which inspired § 3, follow from a reinvestigation of the Schweder–Spjøtvoll–Storey estimator of π_0 (Schweder & Spjøtvoll, 1982; Storey, 2002). The estimator depends on a tuning parameter in $(0,1)$ that is usually denoted by λ. For practical reasons we will write the estimator in terms of $t = 1 - \lambda$. The estimator is

$$\hat{\pi}_0' = \frac{|\{1 \leqslant i \leqslant m : p_i > \lambda\}|}{m(1-\lambda)} = \frac{|\{1 \leqslant i \leqslant m : p_i > 1 - t\}|}{mt}. \tag{1}$$

The heuristics behind the estimate (1) are as follows. The nonnull p-values, i.e., the p-values corresponding to false hypotheses, tend to be smaller than $1-t$, so most of the p-values larger than $1-t$ are null p-values. Since for point null hypotheses the null p-values are standard uniform, one expects approximately $t \times 100\%$ of the null p-values to be larger than $1-t$. Hence, a conservative estimate of the number of null p-values is $t^{-1}|\{i : p_i > 1-t\}|$. Thus, $\hat{\pi}_0'$ is an estimate of π_0. Storey's estimator is related to the concept of accumulation functions, used to estimate false discovery proportions (Li & Barber, 2017; Lei et al., 2021).

We remark that $\hat{\pi}_0'$ can be greater than 1. Consequently, researchers often use $\hat{\pi}_0 = \hat{\pi}_0' \wedge 1$. This estimate is usually no longer biased upwards, but rather biased downwards for large values of π_0, in particular $\pi_0 = 1$.

2.3. *Median unbiased estimation of V and π_0*

Here we derive estimators of V and π_0 that are inspired by the Schweder–Spjøtvoll–Storey estimator. We make the following assumption.

Assumption 1. The following holds:

$$\mathrm{pr}\big\{\big|\{1 \leqslant i \leqslant N : q_i \leqslant t\}\big| > \big|\{1 \leqslant i \leqslant m : p_i \geqslant 1-t\}\big|\big\} \leqslant 0.5. \tag{2}$$

Assumption 1 says that the number of small null p-values, i.e., those less than or equal to t, tends to be smaller than the number of large p-values, i.e., those greater than or equal to $1-t$, both null and nonnull. This assumption, (2), is satisfied in particular if

$$\mathrm{pr}\big\{\big|\{1 \leqslant i \leqslant N : q_i \leqslant t\}\big| > \big|\{1 \leqslant i \leqslant N : q_i \geqslant 1-t\}\big|\big\} \leqslant 0.5. \tag{3}$$

Further, the probability in (3) is equal to

$$\mathrm{pr}\big\{\big|\{1 \leqslant i \leqslant N : q_i \leqslant t\}\big| > \big|\{1 \leqslant i \leqslant N : 1-q_i \leqslant t\}\big|\big\}. \tag{4}$$

If the null p-values $q_1, \ldots, q_N$ are independent and standard uniform, then Assumption 1 is clearly satisfied. As another example, suppose $q = (q_1, \ldots, q_N)$ is symmetric about $1/2$, i.e.,

$$(q_1, \ldots, q_N) \overset{\mathrm{d}}{=} (1-q_1, \ldots, 1-q_N). \tag{5}$$

Then property (4) and hence Assumption 1 also hold. The symmetry property (5) holds for instance if $q_1, \ldots, q_N$ are left- or right-sided p-values from Z-tests based on test statistics $Z_1, \ldots, Z_m$ with joint $N(0, \Sigma)$ distribution. Further, the presence of null p-values that are stochastically larger than uniform, or the presence of many nonnulls, makes it easier for Assumption 1 to be satisfied.

If t is used as a rejection threshold, the number of false positive findings is

$$V(t) = \big|\{1 \leqslant i \leqslant N : q_i \leqslant t\}\big|.$$

Under Assumption 1, with probability at least 0.5 we have

$$V(t) \leqslant \bar{V}(t) = \big|\{1 \leqslant i \leqslant m : p_i \geqslant 1-t\}\big|. \tag{6}$$

In other words, $\bar{V}(t)$ is a 50% confidence upper bound for $V(t)$. We will refer to such bounds as median unbiased estimators for brevity, although writing 'not downward biased' instead of 'unbiased' would be more precise.

This result also leads to a median unbiased estimator of π_0. Indeed, if $V \leqslant \bar{V}$, then $\mathcal{R}$ contains at least $R - \bar{V}$ false hypotheses, so that π_0 is at most

$$\frac{m - R + \bar{V}}{m} = \frac{m - |\{1 \leqslant i \leqslant m : p_i \leqslant t\}| + |\{1 \leqslant i \leqslant m : p_i \geqslant 1 - t\}|}{m}.$$

A reformulation gives the following result.

THEOREM 1. *Suppose that Assumption 1 is satisfied. Then $\bar{V}(t)$, defined in (6), is a median unbiased estimate of $V(t)$. As a consequence, $\bar{\pi}_0 = \bar{\pi}_0' \wedge 1$, where*

$$\bar{\pi}_0' = \frac{|\{1 \leqslant i \leqslant m : p_i > t\}| + |\{1 \leqslant i \leqslant m : p_i \geqslant 1 - t\}|}{m},$$

is a median unbiased estimate of π_0. Further, if $t = 0.5$ and no p-value equals t, then $\bar{\pi}_0'$ is equal to the Schweder–Spjøtvoll–Storey estimate $\hat{\pi}_0'$.

Thus, if the p-values are continuous and $t = 0.5$, then $\bar{\pi}_0' = \hat{\pi}_0'$ with probability 1. For other values of λ, we obtain a median unbiased estimate $\bar{\pi}_0'$ that is slightly different from $\hat{\pi}_0'$. In the Supplementary Material, we provide a theoretical comparison of $E(\bar{\pi}_0')$ with $E(\hat{\pi}_0')$ and obtain the estimate $\bar{\pi}_0'$ in an alternative way; in doing so, we discover a broader class of π_0 estimators.

We write $\bar{\pi}_0 = \min(\bar{\pi}_0', 1)$. In Example 1 and the corresponding Fig. 1, the Schweder–Spjøtvoll–Storey method is applied to 500 simulated p-values.

Example 1 (*Running example, part 1: estimating π_0 and V*). As a toy example we generated 500 independent p-values, 400 of which were uniformly distributed on $[0, 1]$ and 100 of which were stochastically smaller than uniform on $[0, 1]$. Thus we can say that $N = 400$. A scatterplot of the sorted p-values is shown in Fig. 1, as well as a visual illustration of how Storey's estimate $\hat{\pi}_0 m$ of the number of true hypotheses is computed, in the case where $\lambda = 1 - t = 0.8$. Often λ is taken to be smaller, but considering small t instead will turn out to be useful. In this example, Storey's estimate $\hat{\pi}_0 m$ was 410 and our estimate, which is less easy to visualize, was $\bar{\pi}_0 m = 402$. Thus the estimates were close, as is often the case. Since property (5) and hence Assumption 1 are satisfied, we know that $\bar{\pi}_0$ is a median unbiased estimator of π_0. In particular, we know with 50% confidence that there are at least $500 - 402 = 98$ false hypotheses in total.

As explained in this section, we can make this statement stronger by observing that $R(t) = 180$ and $\bar{V}(t) = 82$. The latter means that we know with 50% confidence that there are at least $180 - 82 = 98$ false hypotheses among the hypotheses with p-values below $t = 0.2$.

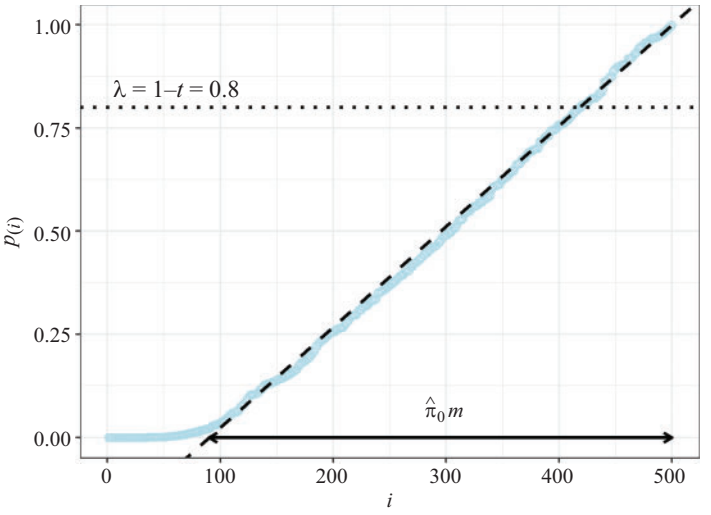

Fig. 1. Illustration of the computation of the Schweder–Spjøtvoll–Storey estimate $\hat{\pi}_0$, based on 500 sorted simulated p-values. The dashed straight line is constructed in such a way that it goes through both (500, 1) and the point where the dotted line intersects the curve of p-values, roughly speaking.

2.4. *Median unbiased estimation of the false discovery proportion*

Define the FDP to be the proportion of false positives,

$$\text{FDP} = \frac{V}{R}, \quad \text{FDP}(t) = \frac{V(t)}{R(t)},$$

which is understood to be 0 when $R = 0$. The median unbiased estimate $\bar{V}$ immediately implies a median unbiased estimate of the FDP.

THEOREM 2. *Suppose that Assumption 1 is satisfied. The variable* $\overline{\text{FDP}}(t) = \bar{V}(t)/R(t)$ *is a median unbiased estimator for the* FDP*, that is,*

$$\text{pr}\{\text{FDP}(t) \leqslant \overline{\text{FDP}}(t)\} \geqslant 0.5. \tag{7}$$

To prove this, we only need to observe that if $V \leqslant \bar{V}$, then $\text{FDP} \leqslant \overline{\text{FDP}}$.

3. CONTROLLING THE MFDP

3.1. *Overview of our method and comparison with false discovery rate control*

In § 2.4 we considered a fixed rejection threshold t and provided a median unbiased estimate for FDP(t). In many situations, one would like to adapt the threshold t based on the data, in such a way that one still obtains a valid median unbiased estimate. Naively choosing t in such a way that an attractive, low estimate of the FDP is obtained can invalidate the procedure, in the sense that inequality (7) no longer holds. In § 3.3, however, we derive a method that provides median unbiased bounds for a large range of t, in such a way that with probability at least 0.5 the bounds are simultaneously valid for all t.

Specifically, we let the user choose some range $\mathbb{T} \subseteq [0, 1]$ of rejection thresholds t of interest before looking at the data. Usually a good choice for $\mathbb{T}$ will be $[0, 1/2]$ or another interval starting at 0. Then we provide 50% confidence upper bounds $B(t)$ for $V(t)$ that are

simultaneously valid over all $t \in \mathbb{T}$:

$$\text{pr}\left[\bigcap_{t\in\mathbb{T}}\{V(t) \leqslant B(t)\}\right] \geqslant 0.5. \tag{8}$$

It then immediately follows that the $B(t)/R(t)$ $(t \in \mathbb{T})$ are simultaneously valid 50% confidence bounds for $\text{FDP}(t)$:

$$\text{pr}\left[\bigcap_{t\in\mathbb{T}}\{\text{FDP}(t) \leqslant B(t)/R(t)\}\right] \geqslant 0.5.$$

Since the threshold t can be chosen based on the data, it can be selected such that $B(t)/R(t)$ is low. In particular, one can prespecify a value $\gamma \in [0, 1]$, for example $\gamma = 0.05$, and take the threshold $t \in \mathbb{T}$ to be the largest value for which $B(t)/R(t) \leqslant \gamma$, if such a t exists. This means that our method can be used to reject a set of hypotheses in such a way that the median of the FDP is bounded by γ:

$$\text{pr}(\text{FDP} \leqslant \gamma) \geqslant 0.5.$$

In other words, we can control the median of the FDP, which we will call the mFDP. Our method is an example of false discovery exceedance control, but with the added property that γ can be chosen post hoc, as we discuss below. Our notation using γ is in line with, e.g., Romano & Wolf (2007) and Basu et al. (2023).

Our method is related to the popular Benjamini–Hochberg procedure, which ensures that $E(\text{FDP}) \leqslant \gamma$ (Benjamini & Hochberg, 1995). The Benjamini–Hochberg procedure ensures that the mean of the FDP is controlled, while our method ensures that the median of the FDP is controlled. The mean and the median of the FDP can be asymptotically equal in some settings where the dependencies among the p-values are not too strong (Neuvial, 2008; Ditzhaus & Janssen, 2019), but there is no general guarantee that they are similar (Romano & Shaikh, 2006; Schwartzman & Lin, 2011). Especially under strong dependence, $\text{mFDP} \leqslant \gamma$ does not need to imply $E(\text{FDP}) \leqslant \gamma$, while the converse does hold in many practical situations. Moreover, unlike mFDP control, FDR control always implies weak control of the familywise error rate (Romano et al., 2008, § 6.4). However, before applying any multiple testing method, we could first perform a global test to enforce weak familywise error rate control (Bernhard et al., 2004).

The most important advantage of our method over that of Benjamini–Hochberg is that it provides simultaneous 50% confidence bounds for the FDP. This allows simultaneous as well as post hoc inference, in the sense that $t \in \mathbb{T}$ can be chosen after seeing the data. Further, we can choose multiple values of t and obtain simultaneously valid statements on the FDP. Moreover, we can choose the target FDP γ post hoc. With the Benjamini–Hochberg procedure, such inference is not possible: if one chooses γ after seeing the data, then the Benjamini–Hochberg procedure can become very anti-conservative. This is discussed in § 3.2.

3.2. *The Benjamini–Hochberg procedure is not flexible*

The main advantage of the method that we propose is that it allows the user to choose one or several rejection thresholds or target FDPs after seeing the data. This contrasts our method with the Benjamini–Hochberg procedure. Indeed, when the target FDR α, or γ in

our notation, is chosen based on the data, then the Benjamini–Hochberg procedure no longer guarantees that $E(\text{FDP}) \leqslant \alpha$, conditional on the post hoc chosen α. When testing a single hypothesis, choosing α post hoc is not generally valid either (Hubbard, 2004; Grünwald, 2023). For simulations illustrating that the Benjamini–Hochberg procedure is not valid post hoc, see the Supplementary Material. Another related result is Fig. 5 in Katsevich & Ramdas (2020), which illustrates, based on simulations, that the Benjamini–Hochberg procedure does not have a simultaneous interpretation. We now give some mathematical examples showing that the Benjamini–Hochberg procedure is not valid post hoc.

Suppose all m hypotheses are true and that the p-values are mutually independent and uniformly distributed on $(0, 1]$. The Benjamini–Hochberg procedure provides m adjusted p-values and rejects all hypotheses with adjusted p-values that are at most α. Let $p^{\text{BH}}_{(1)}$ denote the smallest adjusted p-value. It is well known that if $\alpha \in [0, 1]$ is prespecified and all p-values are independent and uniform on $(0, 1]$, then the probability that the Benjamini–Hochberg procedure rejects any hypotheses is exactly α (Goeman & Solari, 2011). The Benjamini–Hochberg procedure rejects any hypotheses if and only if $p^{\text{BH}}_{(1)} \leqslant \alpha$. Thus, $p^{\text{BH}}_{(1)}$ is uniform on $(0, 1]$.

As a simple example of an α chosen post hoc, take $\alpha = p^{\text{BH}}_{(1)}$. We now show that in this case we no longer have $E(\text{FDP}/\alpha) \leqslant 1$. Since $\alpha = p^{\text{BH}}_{(1)}$, we know that α is uniform on $(0, 1]$. By definition of α, there is always at least one rejected hypothesis. Since all hypotheses are true, this means that we always have $\text{FDP} = 1$. Consequently,

$$E(\text{FDP}/\alpha) = E(1/\alpha) = \int_0^1 x^{-1}\,\mathrm{d}x = \log(x)\Big|_0^1 = \infty,$$

i.e., $E(\text{FDP}/\alpha)$ is completely out of control.

Of course, this is an extreme situation, where α can take any value. We now consider a less extreme situation, where we allow α to take only two values, say a_1 and a_2, with $0 < a_1 \leqslant a_2 < 1$. Specifically, we define α to be a_1 if $p^{\text{BH}}_{(1)} \leqslant a_1$, and otherwise $\alpha = a_2$. This mimics the psychology of a researcher who uses a_2 as a default value for α, but takes α to be a_1 if this still leads to at least one rejection. If $p^{\text{BH}}_{(1)} > a_2$, then we reject nothing, so $\text{FDP} = 0$. Thus, with this definition of α, we have

$$\begin{aligned}
E(\text{FDP}/\alpha) &= \text{pr}\big(p^{\text{BH}}_{(1)} \leqslant a_1\big) E\big(\text{FDP}/a_1 \mid p^{\text{BH}}_{(1)} \leqslant a_1\big) \\
&\quad + \text{pr}\big(a_1 < p^{\text{BH}}_{(1)} \leqslant a_2\big) E\big(\text{FDP}/\alpha \mid a_1 < p^{\text{BH}}_{(1)} \leqslant a_2\big) \\
&= a_1 E\big(1/\alpha \mid p^{\text{BH}}_{(1)} \leqslant a_1\big) + (a_2 - a_1) E\big(1/\alpha \mid a_1 < p^{\text{BH}}_{(1)} \leqslant a_2\big) \\
&= a_1/a_1 + (a_2 - a_1)/a_2 = 1 + (a_2 - a_1)/a_2,
\end{aligned}$$

which always exceeds 1, except when $a_1 = a_2$. As an example, take $a_1 = 0.05$ and $a_2 = 0.1$, which are values often used in practice. This defines a rather limited set of allowed values for α. Nevertheless, we find that $E(\text{FDP}/\alpha) = 1.5$, which is already much larger than 1. The reader can check that if we allow α to take more than two values, then $E(\text{FDP}/\alpha)$ can become huge. Indeed, if we allow α to take any value in $(0, 1]$, then $E(\text{FDP}/\alpha)$ can become infinity, as we saw in the previous example where $\alpha = p^{\text{BH}}_{(1)}$.

These examples show that if α depends on the data, then marginally we often have $E(\text{FDP}/\alpha) > 1$. This means in particular that conditional on α taking a certain value, we

do not generally have $E(\text{FDP}) \leqslant \alpha$. The Supplementary Material includes a simulation study that illustrates this point in various other settings.

3.3. *Simultaneous bounds for the false discovery proportion*

Let $\mathbb{N}$ denote the set of natural numbers. We call a function $B : \mathbb{T} \to \mathbb{N}$ a confidence envelope if it satisfies inequality (8) (cf. Hemerik et al., 2019). We restrict ourselves to such 50% confidence envelopes and do not consider, e.g., 95% confidence envelopes. Let $\mathbb{B}$ be a set of maps $\mathbb{T} \to \mathbb{N}$. Assume that $\mathbb{B}$ is monotone, in the sense that for all $B, B' \in \mathbb{B}$, either $B \geqslant B'$ or $B' \geqslant B$. Here $B \geqslant B'$ means that $B(t) \geqslant B'(t)$ for all $t \in \mathbb{T}$. We call $\mathbb{B}$ the family of candidate envelopes (cf. Hemerik et al., 2019).

We will obtain a confidence envelope by choosing the smallest $B \in \mathbb{B}$ for which $B(t) \geqslant \bar{V}(t)$ for all $t \in \mathbb{T}$. We call this envelope $\tilde{B}$:

$$\tilde{B} = \tilde{B}(p) = \min\Big\{B \in \mathbb{B} : \bigcap_{t\in\mathbb{T}} \big\{B(t) \geqslant \bar{V}(t)\big\}\Big\}.$$

If r is a vector containing, say, l_r p-values, then we write $\mathcal{R}(r,t) = \{1 \leqslant i \leqslant l_r : r_i < t\}$ to make the dependence on the p-values explicit. Analogously, we define $V(r,t)$, $\bar{V}(r,t)$ and $\tilde{B}(r)$. We use the convention that $\mathcal{R}(t) = \mathcal{R}(p,t)$, $V(t) = V(p,t)$, $\bar{V}(t) = \bar{V}(p,t)$ and $\tilde{B} = \tilde{B}(p)$.

We require only the following assumption.

Assumption 2. The following holds:

$$\text{pr}\big\{\tilde{B}(p) \geqslant \tilde{B}(1-q)\big\} \geqslant 0.5. \tag{9}$$

Owing to the monotonicity of the set $\mathbb{B}$, we always have either $\tilde{B}(q) < \tilde{B}(1-q)$ or $\tilde{B}(q) \geqslant \tilde{B}(1-q)$. If the latter inequality has the greater probability, then (9) is always satisfied, since $\tilde{B}(p) \geqslant \tilde{B}(q)$. Assumption 2 is a generalization of Assumption 1, in the sense that if $\mathbb{T}$ is equal to the singleton $\{t\}$, then Assumptions 1 and 2 will coincide for most reasonable choices of $\mathbb{B}$, e.g., for $\mathbb{B}$ as in § 3.4.

Assumption 2 always holds if property (5) is satisfied, regardless of our choice of $\mathbb{B}$. Indeed, if (5) holds, we have $\text{pr}\{\tilde{B}(p) \geqslant \tilde{B}(1-q)\} \geqslant \text{pr}\{\tilde{B}(q) \geqslant \tilde{B}(1-q)\} = \text{pr}\{\tilde{B}(1-q) \geqslant \tilde{B}(q)\}$. Since the latter two probabilities are equal, they are both at least 0.5, so Assumption 2 is satisfied. Moreover, property (5) is not necessary for Assumption 2 to hold, as confirmed by our simulations.

Let $[\,\cdot\,]^+$ be the positive-part function. The following theorem states that $\tilde{B}$ provides simultaneously valid 50% confidence bounds.

THEOREM 3. *Suppose that Assumption 2 holds. Then the function $\tilde{B}$ is a confidence envelope, that is,*

$$\text{pr}\Big[\bigcap_{t\in\mathbb{T}}\{V(t) \leqslant \tilde{B}(t)\}\Big] \geqslant 0.5,$$

$$\text{pr}\Big[\bigcap_{t\in\mathbb{T}}\{\text{FDP}(t) \leqslant \tilde{B}(t)/R(t)\}\Big] \geqslant 0.5.$$

In addition, $\tilde{B}' : \mathbb{T} \to \mathbb{N}$ *defined by*

$$\tilde{B}'(t) = R(t) - \max\{[R(l) - \tilde{B}(l)]^+ : l \in \mathbb{T},\ l \leqslant t)\},$$

which satisfies $\tilde{B}' \leqslant \tilde{B}$*, is also a confidence envelope and potentially improves upon* $\tilde{B}$*.*

The proof is presented in the Supplementary Material, but here we give the intuition behind it. First of all, $\bar{V}(t)$ is a 50% confidence bound for $V(t)$, but not simultaneously over all t. The reason is that if multiple events have probability 0.5, then the probability that all events happen is usually smaller than 0.5. For example, if for $t_1, t_2 \in (0, 1)$ we have $\text{pr}\{V(t_j) \leqslant \bar{V}(t_j)\} \geqslant 0.5$ for $j = 1$ and $j = 2$, then we do not generally have $\text{pr}\{V(t_1) \leqslant \bar{V}(t_1)$ and $V(t_2) \leqslant \bar{V}(t_2)\} \geqslant 0.5$. To get a simultaneous bound for $V(t)$, we usually need a stricter requirement; it is not sufficient to simply define $\tilde{B}(t) = \bar{V}(t)$. In the proof, if $\tilde{B}(p) \geqslant \tilde{B}(1 - q)$, then $\tilde{B}(t) \geqslant V(t)$ for all t. It thus follows from Assumption 2 that our $\tilde{B}$ is a confidence envelope. That $\tilde{B}$ is chosen from a fixed, monotone family is not directly used in the proof. However, if $\tilde{B}$ is chosen from such a family, then $\tilde{B}(p) \geqslant \tilde{B}(q)$ and it follows that if (5) holds, then Assumption 2 is satisfied. Thus, that $\tilde{B}$ is chosen from a fixed, monotone family makes Assumption 2 reasonable. It also allows $\tilde{B}$ to be defined as a simple minimum.

In the rest of this subsection, we provide an extension of the bounds $\tilde{B}'(t)$ and a result on admissibility. It turns out that $\tilde{B}'$ coincides with an envelope obtained through a novel closed testing-based procedure, in the sense of Goeman & Solari (2011) and Goeman et al. (2021). This novel procedure provides a 50% confidence bound for the number of true hypotheses in I, for every subset $I \subseteq \{1, \ldots, m\}$. These bounds are all simultaneously valid with probability at least 50%. We denote these bounds by $\bar{B}(I)$.

Theorem 4. *Write* $\mathcal{M} = \{I \subseteq \{1, \ldots, m\} : I \neq \emptyset\}$*. For every* $I \in \mathcal{M}$ *and* $t \in \mathbb{T}$*, define* $R_I(t) = |\mathcal{R}(t) \cap I| = |\{i \in I : p_i \leqslant t\}|$*. Write*

$$\bar{B}(I) = \max\{|A| : \emptyset \neq A \subseteq I \text{ and } \forall t \in \mathbb{T},\ R_A(t) \leqslant \tilde{B}'(t)\}, \tag{10}$$

where the maximum of an empty set is interpreted as 0*.*

Assume $\mathbb{T} \subseteq [0, 1/2)$ *and* $\text{pr}\{\tilde{B}(q) \geqslant \tilde{B}(1 - q)\} \geqslant 0.5$*. Then*

$$\text{pr}\left[\bigcap_{I \in \mathcal{M}} \{|\mathcal{N} \cap I| \leqslant \bar{B}(I)\}\right] \geqslant 0.5,$$

i.e., the $\bar{B}(I)$ *are simultaneous* 50% *confidence bounds for the number of true hypotheses in* I*. In particular, the function* $\mathbb{T} \to \mathbb{N}$ *defined by* $t \mapsto \bar{B}\{\mathcal{R}(t)\}$ *is a confidence envelope.*

If the local tests discussed in the proof of Theorem 4 are admissible, then the method of Theorem 4 is admissible, in the sense of Theorem 3 in Goeman et al. (2021). The local tests will usually be admissible when $\mathbb{B}$ is any reasonable family, such as the family considered in § 3.4. By Theorem 5 below, if the procedure of Theorem 4 is admissible, then the envelope $\tilde{B}'(t)$ from Theorem 3 is also admissible.

Theorem 5. *For every* $t \in \mathbb{T}$*, the bound* $\tilde{B}'(t)$ *from Theorem 3 is equal to the bound* $\bar{B}\{\mathcal{R}(t)\}$ *from Theorem 4. Moreover, if the procedure from Theorem 4 that provides bounds for all* $I \in \mathcal{M}$ *is admissible, then the envelope* $\tilde{B}'$ *is also admissible. Here admissibility of* $\tilde{B}'$ *means that*

there exists no envelope $B : \mathbb{T} \to \mathbb{N}$ *such that* $B(t) \leqslant \tilde{B}'(t)$ *for all* $t \in \mathbb{T}$ *and such that* $\mathrm{pr}\{\exists\, t \in \mathbb{T} : B(t) < \tilde{B}'(t)\} > 0$.

The admissibility property of our method contrasts with the Benjamini–Hochberg procedure. The latter method is not admissible, since it is uniformly improved upon by the method of Solari & Goeman (2017), for which admissibility is not known. In the rest of this article we will focus on bounds for rejected sets of the form $\mathcal{R}(t) = \{1 \leqslant i \leqslant m : p_i \leqslant t\}$, as constructed in Theorem 3.

3.4. *A default* MFDP *envelope*

The envelope $\tilde{B}$ depends on a general family $\mathbb{B}$ of candidate confidence bounds. The choice of this family can have a large influence on the bounds obtained (cf. Hemerik et al., 2019). An important question is therefore how to choose this set $\mathbb{B}$ in a suitable way. Typically we want $\mathbb{B}$ to contain at least one function B that is a tight upper envelope of the function $t \mapsto \bar{V}(t)$. Between $t = 0$ and, say, $t = 0.5$, the function $\bar{V}(t)$ tends to be roughly linear in t, at least under independence. Thus, it can make sense to also take the candidate envelopes $B \in \mathbb{B}$ to be roughly linear. Also, giving them a small positive intercept will often be useful to avoid having $\tilde{B}$ be too sensitive to p-values near 1.

Further, it is usually appropriate to take $\mathbb{T} = [s_1, s_2]$, where $s_1 \geqslant 0$ is the smallest threshold of interest and $s_2 < 1$ is the largest threshold of interest. Based on these considerations, we propose to use the following default family $\mathbb{B}$ of candidate functions:

$$\mathbb{B} = \{B^\kappa : \kappa \in (0, \infty]\}, \tag{11}$$

with

$$B^\kappa(t) = \big|\{1 \leqslant i \leqslant m : i\kappa - c \leqslant t\}\big| = \left\lfloor \frac{t + c}{\kappa} \right\rfloor.$$

Here, $c \geqslant 0$ is a prespecified small constant. The discrete function B^κ is roughly linear in t and has slope $1/\kappa$.

The choice of c influences the intercept of B^κ and hence the slope and intercept of the resulting envelope $\tilde{B}$. Taking c to be 0 or very small tends to give tighter bounds $\tilde{B}(t)$ for very small t, while taking c a bit larger tends to yield tighter bounds for larger t. We found in simulations that taking $c = 1/(2m)$ usually gave good overall power.

If we take $\mathbb{B}$ as in expression (11), then the confidence envelope becomes

$$\tilde{B} = B^{\kappa_{\max}}, \quad \kappa_{\max} = \max\Big\{\kappa \in (0, \infty] : \bigcap_{t \in \mathbb{T}} \big\{B^\kappa(t) \geqslant \bar{V}(t)\big\}\Big\}. \tag{12}$$

For computational implementation of this method, a useful equivalent formulation is the following, if $\mathbb{T}$ is an interval.

PROPOSITION 1. *Suppose* $\mathbb{T}$ *is of the form* $[s_1, s_2]$ *with* $0 \leqslant s_1 < s_2 \leqslant 1$. *We then have*

$$\kappa_{\max} = \kappa_0 \wedge \min\big\{\kappa_i : 1 \leqslant i \leqslant m,\ 1 - p_i \in \mathbb{T}\big\}, \tag{13}$$

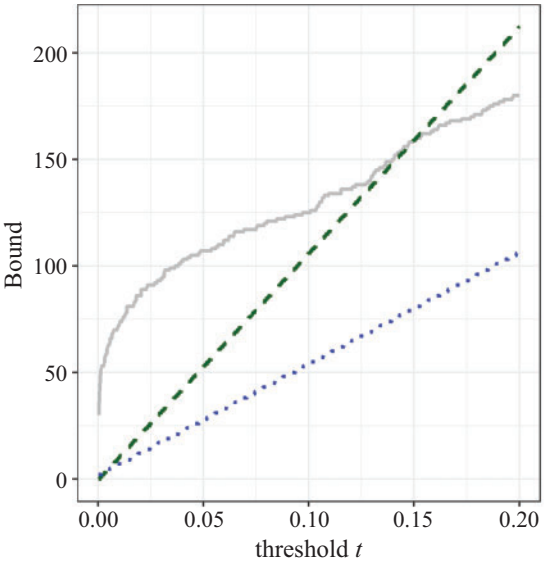

Fig. 2. Graph showing the number of rejections (grey solid) and two confidence envelopes $\tilde{B}$ (green dashed for $c = 0$ and blue dotted for $c = 0.004$) as functions of the rejection threshold t for the running example. The confidence envelopes $\tilde{B}(t)$ are simultaneous 50% confidence upper bounds for the number of false positives $V(t)$, and the intercept and slope depend on the user-specified constant c: for $c = 0$ the intercept is slightly smaller than for $c = 0.004$; indeed, the intercepts are 0 and 2, respectively.

where

$$\kappa_0 = \frac{s_1 + c}{\bar{V}(s_1)} = \frac{s_1 + c}{|\{1 \leqslant j \leqslant m : \ p_j \geqslant 1 - s_1\}|},$$
$$\kappa_i = \frac{1 - p_i + c}{\bar{V}(1 - p_i)} = \frac{1 - p_i + c}{|\{1 \leqslant j \leqslant m : \ p_j \geqslant p_i\}|} \quad (1 \leqslant i \leqslant m).$$

If the denominator is zero, the expression is interpreted as ∞.

We can sometimes straightforwardly improve the envelope $B^{\kappa_{\max}}$ by using the second part of Theorem 3. In Example 2 we continue the running example and compute simultaneous mFDP bounds. Figure 2 shows the confidence envelope and Fig. 3 illustrates how the envelope was determined.

Example 2 (*Running example, part 2: confidence envelopes*). We continue Example 1 by computing confidence envelopes, i.e., simultaneous 50% confidence upper bounds for $V(t)$, the number of false positives, which depends on the threshold t. We take $\mathbb{T} = [0, 0.2]$ and define $\tilde{B}$ as in (12). We compute $\tilde{B}$ for both $c = 0$ and $c = 2/m = 0.004$. These choices for c are somewhat arbitrary. The number of rejections $R(t)$ and the bounds $\tilde{B}(t)$ for both values of c are plotted in Fig. 2. The construction of the confidence envelopes $\tilde{B}$ is illustrated in Fig. 3.

Figure 2 shows that, as expected, near $t = 0$ the number of rejections increases quickly with t. The reason is that there were many p-values near 0, as seen in Fig. 1. By definition (12), the bounds $\tilde{B}(t)$ are roughly linear in t, which can be seen in the figures. We also see that for this specific dataset, the bound $\tilde{B}(t)$ depends strongly on c: it is lower for $c = 0.004$ than for $c = 0$ if t is close to 0, but much higher otherwise. For most values of $t \in [0, 1]$

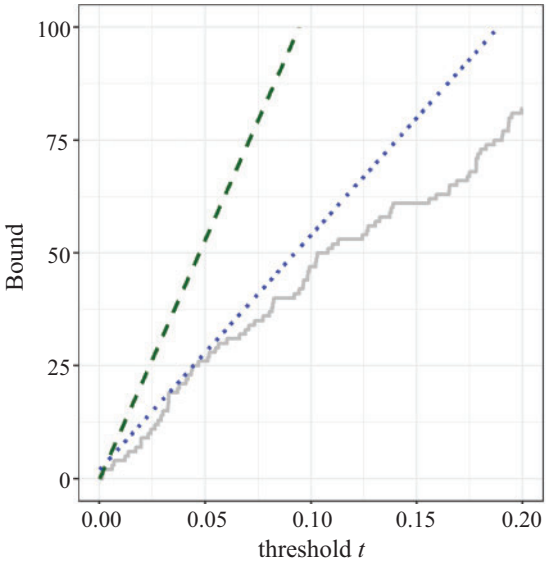

Fig. 3. Illustration of the construction of the confidence envelope for the running example. For every rejection threshold t, $\bar{V}(t)$ (grey solid) is a 50% confidence upper bound for the number of false positives, $V(t)$. The confidence envelope $\tilde{B}(t)$ (green dashed for $c = 0$ and blue dotted for $c = 0.004$) is constructed in such a way that it lies above the pointwise bound $\bar{V}(t)$ for all $t \in \mathbb{T}$. Owing to this construction, the bounds $\tilde{B}(t)$ are simultaneous 50% confidence bounds for $V(t)$. The intercept and slope of $\tilde{B}$ are influenced by the choice of c.

the envelope for $c = 0.004$ is better, i.e., lower, than the envelope for $c = 0$. On the other hand, the smallest cut-offs are often most relevant. Finally, we remark that the bounds in the figures can be somewhat improved by using the last part of Theorem 3. This improvement was used to obtain Fig. 4, where simultaneous 50% confidence bounds for FDP(t) are shown.

3.5. *Controlling the median of the false discovery proportion*

Consider $\gamma \in [0, 1]$. As discussed in § 3.1, we can use any confidence envelope B to guarantee that $\text{pr}(\text{FDP} \leqslant \gamma) \geqslant 0.5$. In other words, we can control the mFDP. By mFDP we mean the median of the distribution that the FDP has, conditional on the data and conditional on γ, which can be chosen after seeing the data. This is stated in the following theorem. The maximum of an empty set is taken to be 0.

THEOREM 6. *Let $B : \mathbb{T} \to \mathbb{N}$ be a confidence envelope, such as $\tilde{B}$. Let the target* FDP *$\gamma \in [0, 1]$ be freely chosen based on the data. Define*

$$t_{\max} = t_{\max}(B, \gamma) = \max\{p_i : \exists\, t \in \mathbb{T} \cap [p_i, 1],\ B(t)/R(t) \leqslant \gamma\}.$$

Reject all hypotheses with p-values at most $t_{\max}$ and denote the FDP *by* FDP_γ. *Then with probability* 0.5 *the* FDP *is at most γ, i.e.,*

$$\text{pr}\{\text{FDP}_\gamma \leqslant \gamma)\} \geqslant 0.5. \tag{14}$$

In fact,

$$\text{pr}\left(\bigcap_{\gamma \in [0,1]} \text{FDP}_\gamma \leqslant \gamma\right) \geqslant 0.5, \tag{15}$$

i.e., the procedure provides mFDP *control simultaneously over all $\gamma \in [0, 1]$.*

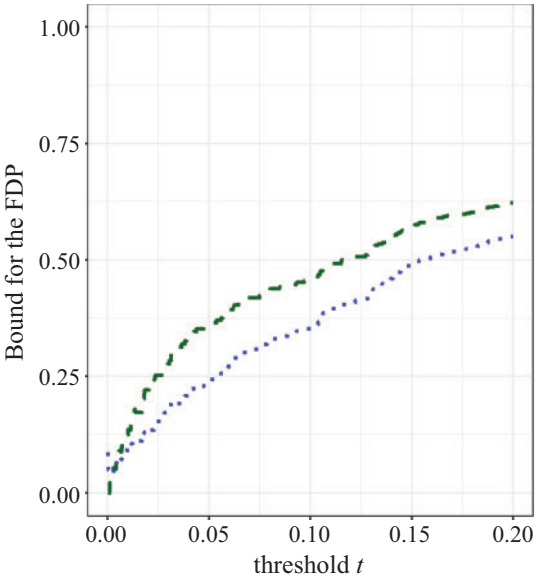

Fig. 4. Simultaneous 50% confidence upper bounds for FDP(t), defined by $\overline{\text{FDP}}(t) = \tilde{B}(t)/R(t)$, for two values of c (green dashed for $c = 0$ and blue dotted for $c = 0.004$). If $c = 0.004$, the bound is greater than zero at $t = 0$; the reason is that $\tilde{B}(0) > 0$ for this value of c. Roughly speaking, the bound $\overline{\text{FDP}}(t)$ then decreases for a while before it starts to increase. If $c = 0$, the bound starts at zero and increases from there.

In other words, if we reject all hypotheses with p-values that are at most $t_{\max}$, then a median unbiased estimate of the FDP is γ. This follows directly from the fact that the estimates $\overline{\text{FDP}}(t)$ ($t \in \mathbb{T}$) are simultaneously valid 50% confidence upper bounds by inequality (8). Inequality (14) holds despite the fact that γ can depend on the data. In fact, with probability at least 50%, $\text{FDP}_\gamma \leqslant \gamma$ simultaneously over all $\gamma \in [0, 1]$. This contrasts our method with many other procedures, which require considering only one rejection criterion, which moreover needs to be chosen in advance (Benjamini & Hochberg, 1995; van der Laan et al., 2004; Lehmann & Romano, 2005; Guo & Romano, 2007; Romano & Wolf, 2007; Neuvial, 2008; Roquain, 2011; Guo et al., 2014; Delattre & Roquain, 2015; Ditzhaus & Janssen, 2019; Döhler & Roquain, 2020; Basu et al., 2023; Miecznikowski & Wang, 2023). In Example 3 we continue the running example and apply our mFDP control method.

Example 3 (*Running example, part 3: controlling the* mFDP). Continuing Example 2, take $\gamma = 0.05$ and consider the confidence envelope $\tilde{B}$ discussed in Example 2. To find a rejection threshold $t_{\max}$ for which we can ensure $\text{mFDP} \leqslant \gamma$, we use Theorem 6. It computes $t_{\max}$ as the largest t for which the estimate in Fig. 4 is at most γ.

Recall that in Example 2 we computed bounds $\tilde{B}(t)$ for both $c = 0$ and $c = 0.004$. For $c = 0$, we now find $t_{\max} = 0.002709$, which is the 54th smallest p-value. Thus, we can reject 54 hypotheses. More precisely, if we reject the 54 smallest p-values, we know that the mFDP is below $\gamma = 0.05$. We remark that $t_{\max}$ is about 27 times higher than the Bonferroni threshold $0.05/500 = 0.0001$.

If $c = 0.004$ then $t_{\max} = 0.001660$, so that we can only reject 53 hypotheses. The reason why $t_{\max}$ is lower if $c = 0.004$ is that for small values of t, the bound $\tilde{B}(t)$ is higher for $c = 0.004$ than for $c = 0$, as can be seen in Fig. 2.

One is allowed to change γ after looking at the data. For instance, if we decrease γ to 0.01, we reject 44 hypotheses for $c = 0$ and reject no hypotheses for $c = 0.004$.

3.6. *Adjusted p-values for* mFDP *control*

Adjusted p-values can be a useful tool in multiple testing. They are defined as the smallest level, for example the smallest γ, at which the multiple testing procedure would reject the hypothesis. Adjusted p-values can be problematic in, for example, FDR control and our context. The reason is that the adjusted p-value does not have an independent meaning and can easily be misinterpreted when taken out of context (Goeman & Solari, 2014, § 5.4). Moreover, an mFDP-adjusted p-value could be 0, which also shows that the interpretation is very different than for real p-values, which cannot be 0. Nevertheless, in our context, adjusted p-values are quite useful, because, once computed, they allow one to check quickly which hypotheses are rejected for various values of γ.

Let B be a confidence envelope and let $1 \leqslant i \leqslant m$. As discussed in § 3.5, B defines an mFDP-controlling procedure. The mFDP-adjusted p-value for H_i is the largest $\gamma \in [0, 1]$ for which H_i is still rejected by the mFDP-controlling procedure. Consequently, if we reject all hypotheses H_i with $p_i^{\mathrm{ad}} \leqslant \gamma$, then $\mathrm{mFDP} \leqslant \gamma$.

PROPOSITION 2. *Let $1 \leqslant i \leqslant m$. Then the value*

$$p_i^{\mathrm{ad}} = \min\{B(t)/R(t) : t \in \mathbb{T} \cap [p_i, 1]\} \tag{16}$$

is an mFDP*-adjusted p-value for H_i, i.e., if we reject all hypotheses H_i with $p_i^{\mathrm{ad}} \leqslant \gamma$, then* $\mathrm{pr}(\mathrm{FDP}_\gamma \leqslant \gamma) \geqslant 0.5$. *Here γ may be chosen based on the data. In fact, inequality* (15) *holds. We take the minimum of an empty set to be ∞.*

Suppose that $\mathbb{T}$, the set of rejection thresholds of interest, is of the form $[s_1, s_2]$. Then we have the following useful reformulation of Proposition 2.

PROPOSITION 3. *Suppose that $\mathbb{T}$ is of the form $[s_1, s_2]$ with $0 \leqslant s_1 < s_2 \leqslant 1$. For each $1 \leqslant i \leqslant m$ with $p_i \leqslant s_2$, the adjusted p-value defined above is then*

$$p_i^{\mathrm{ad}} = \min\big\{B(t)/R(t) : t \in \big[\max\{s_1, p_i\}, s_2\big] \cap \{s_1, p_1, p_2, \ldots, p_m\}\big\}.$$

Given the data, the adjusted p-value is a nondecreasing function of the unadjusted p-value. As a consequence of this and Proposition 3, if $\mathbb{T}$ is of the form $[s_1, s_2]$, Algorithm 1 can be used to efficiently compute the mFDP-adjusted p-values. The algorithm takes the m sorted p-values $p_{(1)}, \ldots, p_{(m)}$ as input and returns the corresponding sorted adjusted p-values.

The idea of the algorithm is to start with computing the largest adjusted p-value(s), continue with the second largest one and so on. The algorithm also uses the fact that if $p_{(i)} > s_2$, then $p_{(i)}^{\mathrm{ad}} = \infty$. It further uses the fact that all hypotheses with unadjusted p-values below s_1 have the same adjusted p-value. Adjusted p-values can be easily computed using the `R` package `mFDP`.

Algorithm 1. Algorithm for computing the mFDP adjusted p-values if $\mathbb{T} = [s_1, s_2]$.

$r \leftarrow |\{1 \leqslant i \leqslant m : p_i \leqslant s_2\}|$.
if $r < m$ then
  $p_{(r+1)}^{\mathrm{ad}}, \ldots, p_{(m)}^{\mathrm{ad}} \leftarrow \infty$.

if $r > 0$ then
  if $s_1 \leqslant p_{(r)}$ then
    $p^{\mathrm{ad}}_{(r)} \leftarrow B(p_{(r)})/R(p_{(r)})$
  else
    $p^{\mathrm{ad}}_{(r)} \leftarrow B(s_1)/R(s_1)$
  $l \leftarrow r-1$
  while $l > 0$ AND $p_{(l)} \geqslant s_1$ do
    $p^{\mathrm{ad}}_{(l)} = \min\{p^{\mathrm{ad}}_{(l+1)}, B(p_{(l)})/R(p_{(l)})\}$
    $l \leftarrow l-1$
  if $l > 0$ then
    $p^{\mathrm{ad}}_{(1)}, \ldots, p^{\mathrm{ad}}_{(l)} \leftarrow \min\{p^{\mathrm{ad}}_{(l+1)}, B(s_1)/R(s_1)\}$
return $p^{\mathrm{ad}}_{(1)}, \ldots, p^{\mathrm{ad}}_{(m)}$

4. Simulations

We performed simulations to assess the error control, power and speed of our method, using R version 4.3 (R Development Core Team, 2024). We compared our approach with three existing methods. The first is the method of Goeman et al. (2019), which exploits the Simes inequality and closed testing and has proven admissibility, like our method. That method is a special case of the one in Goeman & Solari (2011), but Goeman et al. (2019) describes a faster algorithm, although its computational complexity is still not linear like the method proposed here. If one takes $\alpha = 0.5$ in that method, then it provides flexible mFDP control, just like our proposed method. The second method is the procedure for simultaneous FDP control of Katsevich & Ramdas (2020). Taking $\alpha = 0.5$ in that method gives flexible mFDP control, although the method does assume independence. Moreover, that method has proven validity only for $\alpha \leqslant 0.31$, although the authors remark that it is probably also valid for larger α. Finally, we compare our method with the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001), which is the most popular method related to FDP control, although it does not control the mFDP, but rather the FDR. Moreover, it requires choosing γ, called α in Benjamini & Hochberg (1995), before seeing the data. We also consider two so-called adaptive FDR methods, which use an estimate of π_0 to gain power.

In the simulations we considered $m = 10^3$ or 10^4 hypotheses. The p-values were based on Z statistics, computed from simulated data with various dependence structures. The p-values were two-sided unless stated otherwise. To create signal, a number Δ was added to the first $(1-\pi_0)/m$ test statistics. The following dependence structures of the test statistics were considered:

(i) independence, referred to as IN;
(ii) homogeneous positive correlations $\rho = 0.5$, referred to as HO;
(iii) five independent blocks, with positive dependence $\rho = 0.8$ within blocks, referred to as BL;
(iv) 50 negatively dependent blocks with correlations -0.01 and with correlation 0.5 within blocks; the p-values were right-sided so that they were negatively correlated between blocks, referred to as NE.

Further, we varied m, π_0 and the signal Δ.

Table 1. *The error rate of our procedure in various settings with* $m = 10^3$. *The final column shows the simulation-based estimate of the probability that there is a* $0 < \gamma < 1$ *for which* FDP_γ *exceeds* γ*; this probability should not be greater than* 0.5. *For the settings with* $\pi_0 < 1$, *the signal for the false hypotheses was* $\Delta = 3$

π_0	Setting	ρ	pr(error)
1	IN	0	0.499
1	HO	0.2	0.334
1	HO	0.5	0.266
1	HO	0.9	0.330
1	BL	0.5	0.335
1	BL	0.9	0.351
1	NE	−0.01	0.500
0.95	IN	0	0.498
0.95	HO	0.2	0.336
0.95	HO	0.5	0.266
0.95	HO	0.9	0.327
0.95	BL	0.5	0.338
0.95	BL	0.9	0.343
0.95	NE	−0.01	0.501

We computed $\tilde{B}$ as in § 3.4. We took $\mathbb{T} = [0, 0.1]$, i.e., our bounds and mFDP-adjusted p-values were simultaneously valid with respect to all thresholds t in this interval. We took $c = 1/(2m)$ as recommended in § 3.4.

We first assessed whether our method provided appropriate simultaneous mFDP control. The simulation results are shown in Table 1. For each setting, the table reports the estimate of the probability pr{for some $t \in \mathbb{T}$, $V(t) > \tilde{B}(t)$}, which is identical to the probability that there is a $0 < \gamma < 1$ for which FDP_γ exceeds γ. Each estimate was based on 10^4 repeated simulations.

The table confirms the simultaneous control of our method. It can be seen that the estimated error rate is about 0.5 under independence if $\pi_0 = 1$. Indeed, the true error rate is then exactly 0.5. The reason is that in this case $p = q$ and the equality (5) holds, so that the probability in Assumption 2 is exactly 0.5. We see that for $\pi_0 = 0.95$ the error rate is also approximately 0.5, rather than less. This is because our method is quite adaptive. In the setting with negative dependence, $\pi_0 = 1$ and one-sided p-values, the error rate is also exactly 0.5, again because (5) then holds. In the other cases, the method was also valid.

Next, we assessed the power of our method by comparing it with the power of the methods of Goeman et al. (2019) and Katsevich & Ramdas (2020). The power was defined as the average fraction of the false hypotheses that were rejected. For three values of the target FDP γ we estimated the power for the three methods. The results are shown in Fig. 5, where $m = 10^3$ and $\pi_0 = 0.9$. Overall the method of Katsevich & Ramdas (2020) performed least well among the three, especially for $\gamma = 0.01$. This may partly be due to the +1 in their formula for the bound on the number of false positives. Further, for $\gamma = 0.01$, the method of Goeman et al. (2019) had better power than our proposed method. However, as shown in the Supplementary Material, for $m = 10^4$ and $\pi_0 = 0.9$ our method was better than that of Goeman et al. (2019) overall. Further, for $m = 10^4$ and $\pi_0 = 0.5$, our method was clearly better than both competitors, as shown in Fig. 6. This can be understood by recognizing

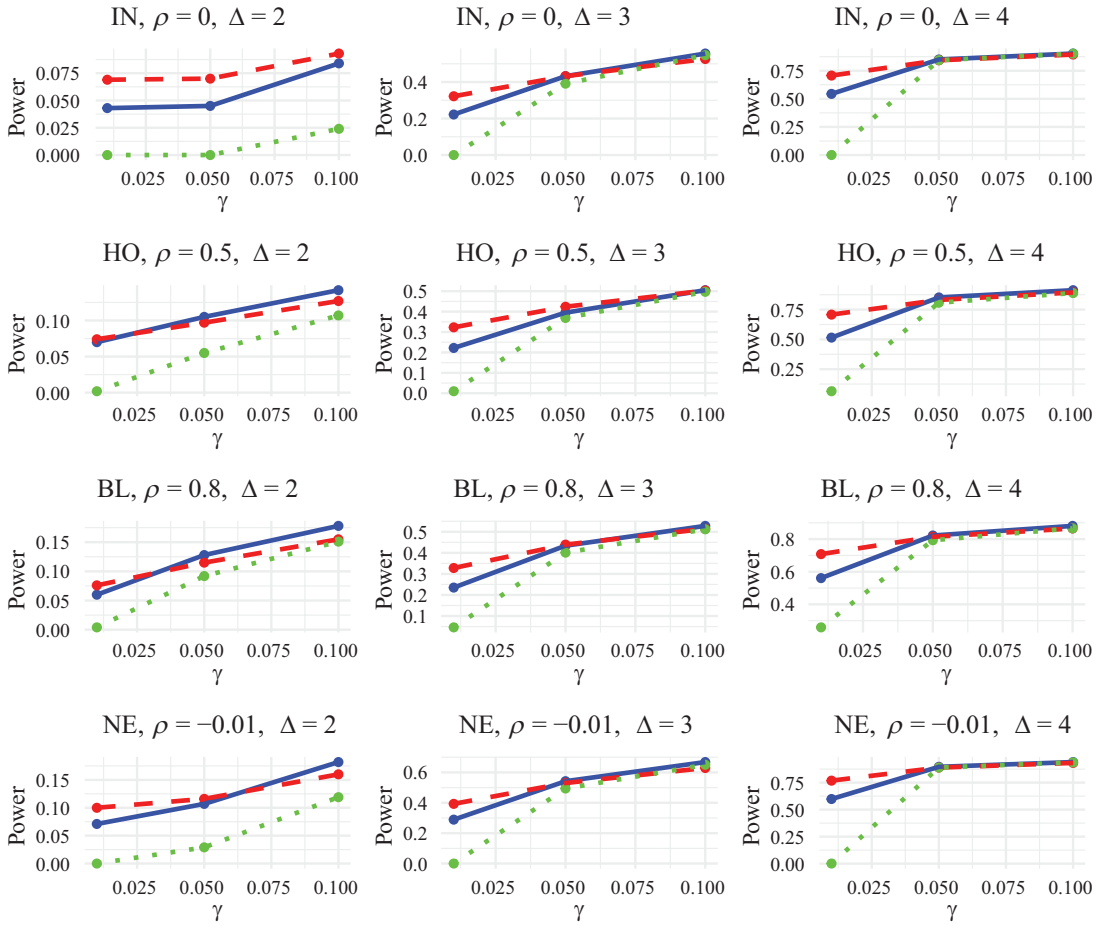

Fig. 5. The power of our proposed method (blue solid) and the methods of Goeman et al. (2019) (red dashed) and Katsevich & Ramdas (2020) (green dotted) plotted against γ for various settings with $m = 10^3$ and $\pi_0 = 0.9$. Each estimate is based on 10^4 simulations.

that the method of Katsevich & Ramdas (2020) is not adaptive, i.e., it is conservative when π_0 is far from 1.

Further, our method was orders of magnitude faster than the method of Goeman et al. (2019), especially for large m. For example, in the setting of the first panel of Fig. 6, our method took 1.7×10^{-2} seconds on average, while that of Goeman et al. (2019) took 4.8 seconds on average. The method of Katsevich & Ramdas (2020) was the fastest, taking 8.6×10^{-4} seconds on average. The reason is that the bounds for $V(t)$ provided by that method depend only on m and t and not on the data.

Finally, for the same simulation settings we computed the power of the Benjamini–Hochberg procedure and two adaptive versions thereof. The results are reported in Table 2. The first column shows the power of the standard Benjamini–Hochberg procedure. The other columns show the power of two versions of the right-boundary procedure of Liang & Nettleton (2012), which makes the Benjamini–Hochberg method more powerful by using an estimate of π_0. The first version, BH*, is their original procedure based on Storey's estimator $\hat{\pi}_0'$. The second one, BH**, is the same, but based on our proposed estimator $\bar{\pi}_0'$. Since the

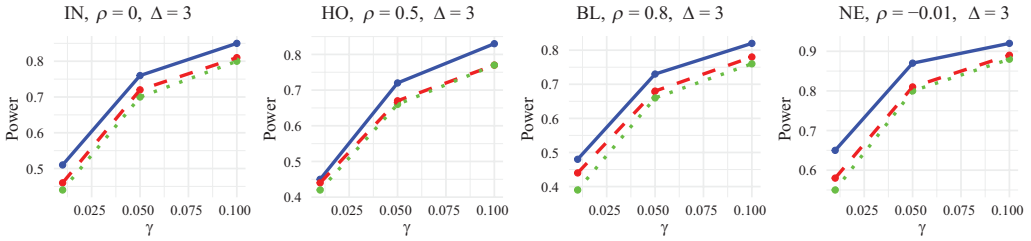

Fig. 6. The power of our proposed method (blue solid) and the methods of Goeman et al. (2019) (red dashed) and Katsevich & Ramdas (2020) (green dotted) plotted against γ for various settings with $m = 10^4$ and $\pi_0 = 0.5$. Each estimate is based on 10^3 simulations.

Table 2. *The power of the Benjamini–Hochberg procedure and two adaptive versions of it. The target* FDR α*, i.e.,* γ*, was* 0.05*; each estimate is based on* 10^4 *simulations*

			Method		
Setting	ρ	Δ	BH	BH*	BH**
IN	0	2	0.058	0.062	0.062
IN	0	3	0.496	0.511	0.512
IN	0	4	0.879	0.886	0.886
HO	0.5	2	0.099	0.125	0.122
HO	0.5	3	0.466	0.494	0.485
HO	0.5	4	0.861	0.910	0.904
BL	0.8	2	0.129	0.153	0.150
BL	0.8	3	0.472	0.503	0.499
BL	0.8	4	0.842	0.863	0.862
NE	−0.01	2	0.120	0.130	0.130
NE	−0.01	3	0.598	0.617	0.617
NE	−0.01	4	0.919	0.925	0.925

Benjamini–Hochberg procedure and its adaptive versions require choosing α beforehand, we show the power only for $\alpha = 0.05$, i.e., $\gamma = 0.05$.

Comparing Fig. 5 and Table 2 shows that for $\gamma = 0.05$, the power of our method was roughly equal to that of the Benjamini–Hochberg procedure, yet often slightly lower. However, our method provides simultaneous bounds and γ can be chosen after seeing the results. As expected, the adaptive Benjamini–Hochberg methods had a bit more power than the Benjamini–Hochberg procedure. The adaptive methods performed similarly to each other. We found that they provided valid FDR control in all the settings, except in HO, where the FDR of BH* varied around 0.08 and the FDR of BH** varied around 0.07.

5. Discussion

This article has introduced an exploratory multiple testing approach, which is useful in particular because the user is allowed to freely choose rejection thresholds based on the data. This is what many researchers would like to do, but which many of the most popular methods do not allow. We have presented a result on the admissibility of our approach, and the simulations demonstrate good power, especially in settings with many false hypotheses. Moreover, the power properties can be influenced by the user, who can select an appropriate family of candidate envelopes $\mathbb{B}$. The choice of the range $\mathbb{T}$ of rejection thresholds also affects the power, since the method focuses power on the thresholds within this range.

Since our method essentially provides estimates for the FDP without confidence intervals, we encourage users to also compute a confidence interval using, for example, the methods listed in § 1. However, as discussed, the methods among those that are valid under dependence have limited power. This means that the confidence interval for the FDP may contain 1, even when there are several strong signals. If permutation of the data is valid, this can often be used to construct tighter confidence intervals (Hemerik et al., 2019; Blain et al., 2022; Andreella et al., 2023).

Our simulations illustrate that for a given γ, the Benjamini–Hochberg procedure tends to have slightly greater power than our method, but our method has the advantage that it provides post hoc inference. Indeed, we have shown that the Benjamini–Hochberg procedure often becomes too liberal when α is chosen post hoc. On the other hand, we control the median of the FDP, which may not always be as appealing as control of the mean. To further illustrate the utility of our method, in the Supplementary Material we provide a data analysis of real RNA-Seq data. Here we further explain how our method's flexibility can lead to additional insights into the data.

Both our proposed method and the Benjamini–Hochberg procedure have certain proven finite-sample, theoretical guarantees, in particular under independence. None of the methods are guaranteed to be valid under an unknown dependence structure. However, there is much evidence that the Benjamini–Hochberg procedure is valid for many dependence structures. Likewise, we did not find a simulation setting in which our method was invalid.

Besides FDP estimators, we have provided a novel π_0 estimator. We have conducted simulations where this estimator was used within an adaptive Benjamini–Hochberg approach. Future work may more extensively assess our estimator in such settings. Further avenues for potential future research are discussed in the Supplementary Material. There we consider more general estimates of π_0 and $V(t)$, which can be combined with the approach in § 3.3 of constructing simultaneous mFDP bounds.

'Uniform' or 'simultaneous' control usually means that the probability of a union of events is kept below some value (Genovese & Wasserman, 2004; Meinshausen, 2006; Blanchard et al., 2020; Goeman et al., 2021). Since FDR control is not defined as controlling a probability, simultaneous FDR control is in that sense undefined. However, interestingly, Corollary 1 in Katsevich & Ramdas (2018) provides what might be called 'simultaneous FDR control', assuming the p-values are independent. In particular, there α can be chosen post hoc while still guaranteeing that the FDR is at most α.

Supplementary material

The Supplementary Material contains additional theory and simulations, an analysis of RNA-Seq data and proofs of Theorems 3–6 and Propositions 1–3.

References

Andreella, A., Hemerik, J., Weeda, W., Finos, L. & Goeman, J. (2023). Permutation-based true discovery proportions for fMRI cluster analysis. *Statist. Med.* **42**, 2311–40.

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–85.

Basu, P., Fu, L., Saretto, A. & Sun, W. (2023). Empirical Bayes control of the false discovery exceedance. *arXiv:* 2111.03885v3.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B **57**, 289–300.

Benjamini, Y. & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* **29**, 1165–88.
Bernhard, G., Klein, M. & Hommel, G. (2004). Global and multiple test procedures using ordered *p*-values—a review. *Statist. Papers* **45**, 1–14.
Blain, A., Thirion, B. & Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage* **260**, 119492.
Blanchard, G., Neuvial, P. & Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *Ann. Statist.* **48**, 1281–303.
Delattre, S. & Roquain, E. (2015). New procedures controlling the false discovery proportion via Romano–Wolf's heuristic. *Ann. Statist.* **43**, 1141–77.
Dickhaus, T. (2014). *Simultaneous Statistical Inference: With Applications in the Life Sciences*. Berlin: Springer.
Ditzhaus, M. & Janssen, A. (2019). Variability and stability of the false discovery proportion. *Electron. J. Statist.* **13**, 882–910.
Döhler, S. & Roquain, E. (2020). Controlling the false discovery exceedance for heterogeneous tests. *Electron. J. Statist.* **14**, 4244–72.
Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Am. Statist. Assoc.* **102**, 93–103.
Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statist. Meth. Med. Res.* **17**, 347–88.
Genovese, C. & Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.* **32**, 1035–61.
Genovese, C. R. & Wasserman, L. (2006). Exceedance control of the false discovery proportion. *J. Am. Statist. Assoc.* **101**, 1408–17.
Goeman, J. J., Hemerik, J. & Solari, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *Ann. Statist.* **49**, 1218–38.
Goeman, J. J., Meijer, R. J., Krebs, T. J. & Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika* **106**, 841–56.
Goeman, J. J. & Solari, A. (2011). Multiple testing for exploratory research. *Statist. Sci.* **26**, 584–97.
Goeman, J. J. & Solari, A. (2014). Multiple hypothesis testing in genomics. *Statist. Med.* **33**, 1946–78.
Grünwald, P. (2023). Beyond Neyman-Pearson. *arXiv:* 2205.00901v2.
Guo, W., He, L. & Sarkar, S. K. (2014). Further results on controlling the false discovery proportion. *Ann. Statist.* **42**, 1070–101.
Guo, W. & Romano, J. (2007). A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statist. Appl. Genet. Molec. Biol.* **6**, 1–33.
Harvey, C. R., Liu, Y. & Saretto, A. (2020). An evaluation of alternative multiple testing methods for finance applications. *Rev. Asset Pricing Stud.* **10**, 199–248.
Hemerik, J. & Goeman, J. J. (2018). False discovery proportion estimation by permutations: Confidence for significance analysis of microarrays. *J. R. Statist. Soc.* B **80**, 137–55.
Hemerik, J., Solari, A. & Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika* **106**, 635–49.
Hoang, A.-T. & Dickhaus, T. (2022). On the usage of randomized *p*-values in the Schweder–Spjøtvoll estimator. *Ann. Inst. Statist. Math.* **74**, 289–319.
Hochberg, Y. & Benjamini, Y. (1990). More powerful procedures for multiple significance testing. *Statist. Med.* **9**, 811–18.
Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between p's and α's in psychological research. *Theory Psychol.* **14**, 295–327.
Katsevich, E. & Ramdas, A. (2018). Towards "simultaneous selective inference": Post hoc bounds on the false discovery proportion. *arXiv:* 1803.06790v3.
Katsevich, E. & Ramdas, A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *Ann. Statist.* **48**, 3465–87.
Langaas, M., Lindqvist, B. H. & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Statist. Soc.* B **67**, 555–72.
Lehmann, E. L. & Romano, J. P. (2005). Generalizations of the familywise error rate. *Ann. Statist.* **33**, 1138–54.
Lei, L. & Fithian, W. (2018). AdaPT. *J. R. Statist. Soc.* B **80**, 649–79.
Lei, L., Ramdas, A. & Fithian, W. (2021). A general interactive framework for false discovery rate control under structural constraints. *Biometrika* **108**, 253–67.
Li, A. & Barber, R. F. (2017). Accumulation tests for FDR control in ordered hypothesis testing. *J. Am. Statist. Assoc.* **112**, 837–49.
Liang, K. & Nettleton, D. (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *J. R. Statist. Soc.* B **74**, 163–82.
Luo, D., He, Y., Emery, K., Noble, W. S. & Keich, U. (2022). Competition-based control of the false discovery proportion. *arXiv:* 2011.11939v3.

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scand. J. Statist.* **33**, 227–37.
Meinshausen, N. & Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.* **34**, 373–93.
Miecznikowski, J. & Wang, J. (2023). Exceedance control of the false discovery proportion via high precision inversion method of Berk-Jones statistics. *Comp. Statist. Data Anal.* **185**, 107758.
Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Statist.* **2**, 1065–110.
R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
Rajchert, A. & Keich, U. (2022). Controlling the false discovery rate via knockoffs: is the +1 needed? *arXiv:* 2204.13248v2.
Rogan, W. J. & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* **107**, 71–6.
Romano, J. P. & Shaikh, A. M. (2006). Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.* **34**, 1850–73.
Romano, J. P., Shaikh, A. M. & Wolf, M. (2008). Formalized data snooping based on generalized error rates. *Economet. Theory* **24**, 404–47.
Romano, J. P. & Wolf, M. (2007). Control of generalized error rates in multiple testing. *Ann. Statist.* **35**, 1378–408.
Roquain, E. (2011). Type I error rate control for testing many hypotheses: A survey with proofs. *arXiv:* 1012.4078v2.
Rosenblatt, J. D. (2021). Prevalence estimation. In *Handbook of Multiple Comparisons*. Boca Raton, Florida: Chapman and Hall/CRC, pp. 183–210.
Schwartzman, A. & Lin, X. (2011). The effect of correlation in false discovery rate estimation. *Biometrika* **98**, 199–214.
Schweder, T. & Spjøtvoll, E. (1982). Plots of *p*-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.
Solari, A. & Goeman, J. J. (2017). Minimally adaptive BH: A tiny but uniform improvement of the procedure of Benjamini and Hochberg. *Biomet. J.* **59**, 776–80.
Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc.* B **64**, 479–98.
van der Laan, M. J., Dudoit, S. & Pollard, K. S. (2004). Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Statist. Appl. Genet. Molec. Biol.* **3**, 15.
Vesely, A., Finos, L. & Goeman, J. J. (2023). Permutation-based true discovery guarantee by sum tests. *J. R. Statist. Soc.* B **64**, 664–83.

[*Received on* 1 *May* 2023. *Editorial decision on* 18 *March* 2024]