New FDR bounds for discrete and heterogeneous tests

Sebastian Döhler, Guillermo Durand, Etienne Roquain

Electronic Journal of Statistics
Vol. 12 (2018) 1867–1900
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1441

New FDR bounds for discrete and heterogeneous tests

Sebastian Döhler

Darmstadt University of Applied Sciences,
D-64295 Darmstadt, Germany
e-mail: sebastian.doehler@h-da.de

Guillermo Durand

Sorbonne Université,
Laboratoire de Probabilités Statistique et Modélisation, LPSM,
F-75005 Paris, France
e-mail: guillermo.durand@upmc.fr

and

Etienne Roquain

Sorbonne Université,
Laboratoire de Probabilités Statistique et Modélisation, LPSM,
F-75005 Paris, France
e-mail: etienne.roquain@upmc.fr

Abstract: To find interesting items in genome-wide association studies or next generation sequencing data, a crucial point is to design powerful false discovery rate (FDR) controlling procedures that suitably combine discrete tests (typically binomial or Fisher tests). In particular, recent research has been striving for appropriate modifications of the classical Benjamini-Hochberg (BH) step-up procedure that accommodate discreteness and heterogeneity of the data. However, despite an important number of attempts, these procedures did not come with theoretical guarantees. In this paper, we provide new FDR bounds that allow us to fill this gap. More specifically, these bounds make it possible to construct BH-type procedures that incorporate the discrete and heterogeneous structure of the data and provably control the FDR for any fixed number of null hypotheses (under independence). Markedly, our FDR controlling methodology also allows to incorporate the quantity of signal in the data (corresponding therefore to a so-called π_0-adaptive procedure) and to recover some prominent results of the literature. The power advantage of the new methods is demonstrated in a numerical experiment and for some appropriate real data sets.

MSC 2010 subject classifications: Primary 62H15; secondary 62Q05.
Keywords and phrases: False discovery rate, heterogeneous data, discrete hypothesis testing, type I error rate control, adaptive procedure, step-up algorithm, step-down algorithm.

Received November 2017.

1. Introduction

Multiple testing procedures are now routinely used to find significant items in massive and complex data. An important focus has been given to methods controlling the false discovery rate (FDR) because this scalable type I error rate "survives" to high dimension. Since the original procedure of Benjamini and Hochberg (1995), much effort has been undertaken to design FDR controlling procedures that adapt to various underlying structures of the data, such as the quantity of signal, the signal strength and the dependencies, among others.

In this work, our motivation is to deal with adaptation to discrete and heterogeneous data. This type of data arises in many relevant applications, in particular when data are represented by counts. Examples can be found in clinical studies (see, e.g., Westfall and Wolfinger, 1997), genome-wide association studies (GWAS) (see, e.g., Dickhaus et al., 2012) and next generation sequencing data (NGS) (see, e.g., Chen and Doerge, 2015b). It is well known (see, e.g., Westfall and Wolfinger, 1997) that using discrete test statistics can generate a severe power loss, already at the stage of the single tests. A consequence is that using "blindly" the BH procedure with discrete p-values will control the FDR in a too conservative manner. Therefore, more powerful procedures that avoid this conservatism are much sought after in applications, see for instance Karp et al. (2016), van den Broek et al. (2015) and Dickhaus et al. (2012).

In the literature, building multiple testing procedures that take into account the discreteness of the test statistics has a long history that can be traced back to Tukey and Mantel (1980). Some null hypotheses can be *a priori* excluded from the study because the corresponding tests are unable to produce sufficiently small p-values. This results in a multiplicity reduction that should increase the power. While this idea has been exploited in Tarone (1990) and in a more general manner in Westfall and Wolfinger (1997) for family-wise error rate, a first attempt was made for FDR in Gilbert (2005). More recently, Heyse (2011) has proposed a more powerful solution, relying on the following averaged cumulative distribution function (c.d.f.):

$$\overline{F}(t) = \frac{1}{m}\sum_{i=1}^{m} F_i(t), \quad t \in [0,1], \tag{1}$$

where each F_i corresponds to the c.d.f. of the i-th test p-value under the null hypothesis. To illustrate the potential benefit of using $\overline{F}$, Figure 1 displays this function for the pharmacovigilance data from Heller and Gur (2011) (see Section 5 for more details). It is important to note that heterogeneity and discreteness structures are both essential in (1): on the one hand, without any heterogeneity (all the F_i's are equal), we have $\overline{F}(t) = F_1(t)$ and there is no benefit in averaging the null; on the other hand, without discreteness, the F_i's are essentially invertible and the p-values can be transformed to be (continuous) uniform under the null, so that the standard BH procedure can be applied. Both structures commonly arise when multiple conditional tests are performed, for which the heterogeneity and discreteness come from marginal counts of contingency tables, e.g., for multiple Fisher exact tests (see simulations in Section 6).

The critical values of the Heyse procedure can be obtained by inverting $\overline{F}$ at the values $\alpha k/m, 1 \leq k \leq m$. Thus, the smaller the $\overline{F}$-values, the larger the critical values. For the example depicted in Figure 1, Heyse critical values improve the BH critical values roughly by a factor 3, thereby yielding a potentially strong rejection enhancement. Furthermore, since the functions F_i's are known in practice, so is $\overline{F}$. Hence, the user has a good prior idea of the improvements reachable by this discrete approach. Unfortunately, the Heyse procedure does not rigorously control the FDR in general; counterexamples are provided in Heller and Gur (2011) and Döhler (2016) (and also in Appendix B.1).

Meanwhile, different solutions have been explored by modifying directly the p-values, either by randomisation (see Habiger, 2015 and references therein), or by shrinking them to build so-called mid p-values (see Heller and Gur, 2011 and references therein). While randomised approaches possess attractive theoretical properties, they are often criticised for their lack of reproducibility (see, e.g., Berger, 1996 and Ripamonti et al., 2017). Other approaches incorporate discreteness and heterogeneity by constructing less conservative FDR estimates, see, e.g., Pounds and Cheng (2006), or by combining grouping and weighting approaches, see Chen and Doerge (2015b).

Overall, although many new procedures have been proposed in the literature, only few of them have been proved to achieve a rigorous FDR control under standard conditions, especially in the finite sample case. To the best of our knowledge, we can only refer to the discretised version of the procedure of Benjamini and Liu (1999) introduced by Heller and Gur (2011) and to the asymptotic work of Ferreira (2007). Our paper offers a solution to this problem by presenting new procedures that achieve both theoretical validity and good practical performance. These procedures are readily implemented in computer software and are therefore easy to apply. Moreover, since neither randomisation nor any additional choice of tuning parameters is necessary, their results are easy to interpret.

The paper is organised as follows: after having precisely defined the setting in Section 2, we introduce in Section 3 new procedures relying on the following modifications of the $\overline{F}$ function:

$$\overline{F}_{\mathrm{SU}}(t) = \frac{1}{m}\sum_{i=1}^{m} \frac{F_i(t)}{1 - F_i(\tau_m)}; \quad \overline{F}_{\mathrm{SD}}(t) = \frac{1}{m}\sum_{i=1}^{m} \frac{F_i(t)}{1 - F_i(t)}, \quad t \in [0,1],$$

(with the convention $1/0 = +\infty$), where an appropriate choice of τ_m is made. To feel how light these modifications are, Figure 1 displays these functions and shows they are very close to the original $\overline{F}$ for small values of t. In addition, we also introduce more powerful "adaptive" versions, meaning that the derived critical values are designed in a way that "implicitly estimates" the overall proportion of true null hypotheses and thus may outperform the original Heyse procedure. Next, in Section 4, we establish rigorous FDR control of the corresponding non-adaptive and adaptive procedures under standard conditions. Our proofs, presented in Section 8, rely on new bounds on FDR that generalise some prominent results of the multiple testing literature. These bounds are the

main mathematical contributions of the paper and are interesting in their own right, beyond the discrete setting. Also, to explore in detail the improvement of our procedures, we analyse both real and simulated data in Sections 5 and 6. Finally, while some additional procedures are presented in Appendix A, other complementary results are provided in Appendices B, C and D.

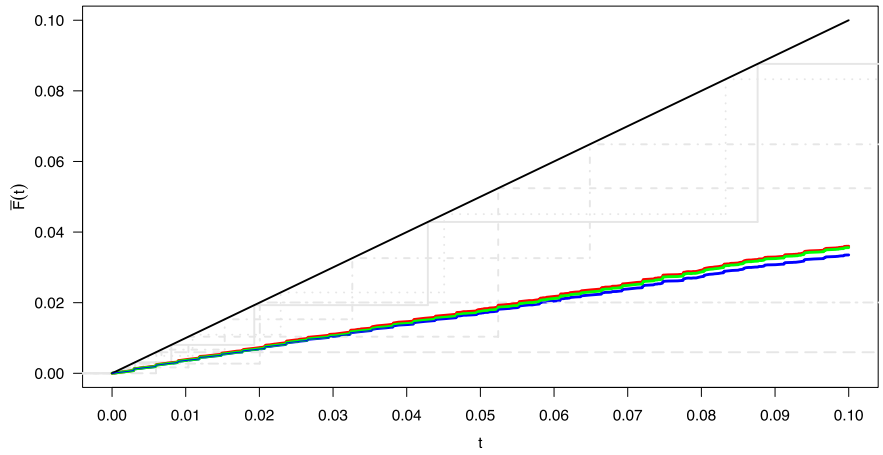


FIG 1. *Plots of variants of* $\overline{F}$ *for the pharmacovigilance data. The solid black line corresponds to the uniform case, the discrete variants are represented by blue (for* $\overline{F}$*), green (for* $\overline{F}_{\mathrm{SD}}$*) and red (for* $\overline{F}_{\mathrm{SU}}$*) solid lines. Some* F_i*'s are displayed in light grey by using different line types.*

2. Preliminaries

2.1. General model

Let us observe a random variable X, defined on a probabilistic space and valued in an observation space $(\mathcal{X}, \mathfrak{X})$. We consider a set $\mathcal{P}$ of possible distributions for the distribution of X and we denote the true one by P. We assume that m null hypotheses $H_{0,i}$, $1 \leq i \leq m$, are available for P and we denote the corresponding set of true null hypotheses by $\mathcal{H}_0(P) = \{1 \leq i \leq m : H_{0,i} \text{ is satisfied by } P\}$. We also denote by $\mathcal{H}_1(P)$ the complement of $\mathcal{H}_0(P)$ in $\{1, \ldots, m\}$ and by $m_0(P) = |\mathcal{H}_0(P)|$ the number of true nulls.

We assume that the user has at hand a set of p-values to test each null, that is, a set of random variables $\{p_i(X), 1 \leq i \leq m\}$, valued in $[0, 1]$. Throughout the paper, we also make the important (but classical) following assumption:

$$\begin{array}{c} \{p_i(X), i \in \mathcal{H}_0\} \text{ consists of independent variables} \\ \text{and is independent of } \{p_i(X), i \in \mathcal{H}_1\}. \end{array} \tag{Indep}$$

Note that (Indep) is satisfied when all the p-values $p_i(X)$, $1 \leq i \leq m$, are mutually independent. Nevertheless, this setting also encompasses situations where there are some dependencies between the p-values under the alternative.

Now, we denote $\mathcal{F} = \{F_i, 1 \le i \le m\}$, where for each $i \in \{1, \dots, m\}$, we let

$$F_i(t) = \sup_{P \in \mathcal{P} \,:\, i \in \mathcal{H}_0(P)} \mathbf{P}_{X \sim P}(p_i(X) \le t), \;\; t \in [0,1], \;\; 1 \le i \le m,$$

which is assumed to be *known*. Note that we necessarily have $F_i(\cdot)$ non decreasing, $F_i(t) \in [0,1]$, $F_i(1) = 1$ and we add the technical condition $F_i(0) = 0$. Loosely, each F_i corresponds to the (least favorable) cumulative distribution function of p_i under the null. Above, we have taken the supremum to cover the case where the null hypothesis is composite: in that situation, each F_i is adjusted according to the least favorable configuration within the null $H_{0,i}$.

Here are some conditions on $\mathcal{F}$ that will be useful to compare some of the studied procedures (these conditions are *not* assumed in our results unless explicitly mentioned):

$$F_i(t) \le t, \;\; t \in [0,1], \;\; 1 \le i \le m, \tag{2}$$

$$F_i(t) = t, \;\; t \in [0,1], \;\; 1 \le i \le m. \tag{3}$$

Condition (2) ensures that the p-values have marginals stochastically lower-bounded by a uniform variable under the null, called a *super-uniform* distribution in the sequel. This is the classical setting which is used in most of the work dealing with FDR controlling theory, see, e.g., Benjamini and Hochberg (1995). Condition (3) is more restrictive: if each null hypothesis is a singleton, it is equivalent to the p-values having uniform marginals under the null.

2.2. Discrete and continuous modelling

In order to describe the overall support of p-value distributions, we assume one of the two following situations to be at hand throughout the paper (except in Section 4 which is written in a more general manner):

- Continuous case: for all $i \in \{1, \dots, m\}$, F_i is continuous. In that case, we let $\mathcal{A}_i = [0,1]$, $1 \le i \le m$ and $\mathcal{A} = \cup_{i=1}^m \mathcal{A}_i = [0,1]$, which corresponds to the overall p-value support.
- Discrete case: each p-value p_i (both under the null and alternative) takes values in some finite set $\mathcal{A}_i$. We denote $\mathcal{A} = \cup_{i=1}^m \mathcal{A}_i$ the overall p-value support.

The continuous setting is typically valid in situations where the p-values are calibrated from test statistics having a continuous distribution under the null. In this situation, (3) is often satisfied. The discrete setting typically arises in situations where the p-values are calibrated from test statistics having a finitely supported distribution under the null. In this situation, we generally have that (3) holds true only on the support of F_i, that is,

$$F_i(t) = t, \;\; t \in \mathcal{A}_i, \;\; 1 \le i \le m. \tag{4}$$

In the discrete framework, let us underline that while (4) will typically hold, the equality $F_i(t) = t$, $t \in \mathcal{A}$ will fail in general because $\mathcal{A}$ contains points of $\mathcal{A}_j$

for $j \neq i$. As a result, $\overline{F}(t)$ defined by (1) will be smaller than t in general (see Figure 1), which is exactly the property that we want to exploit in this paper.

To illustrate the above framework, we provide below two simple examples (for more advanced examples, see for instance Chen and Doerge (2015b)).

Example 1 (Gaussian testing)**.** *Observe $X = (X_i)_{1\leq i\leq m}$ with independent coordinates and marginals $X_i \sim \mathcal{N}(\mu_i, 1)$, where $\mu_i \in \mathbb{R}$ is the parameter of interest, $1 \leq i \leq m$. In that situation, a possible hypothesis testing problem is to consider the nulls $H_{0,i}$: "$\mu_i \leq 0$" against $H_{1,i}$: "$\mu_i > 0$". Then $p_i(X) = 1 - \Phi(X_i)$, $1 \leq i \leq m$, is a family of p-values satisfying (3) (where Φ denotes the c.d.f. of a standard Gaussian variable).*

Example 2 (Binomial testing)**.** *Observe $X = (X_i)_{1\leq i\leq m}$ with independent coordinates and marginals $X_i \sim \mathcal{B}(n_i, \theta_i)$, where $n_i \geq 1$ is known and $\theta_i \in (0, 1)$ is the parameter of interest, $1 \leq i \leq m$. In that situation, a possible hypothesis testing problem is to consider the nulls $H_{0,i}$: "$\theta_i \leq 1/2$" against $H_{1,i}$: "$\theta_i > 1/2$". Then $p_i(X) = T_i(X_i)$, $1 \leq i \leq m$, define a family of p-values where $T_i(x) = 2^{-n_i}\sum_{j=x}^{n_i}\binom{n_i}{j}$ is the upper-tail distribution function of a binomial distribution of parameters $(n_i, 1/2)$. The support of the p-values under the null and alternative is given by the values $2^{-n_i}\sum_{j=K_i-k}^{n_i}\binom{n_i}{j}$, $1 \leq k \leq K_i$, where $K_i = n_i + 1$ and $1 \leq i \leq m$. We easily check in that case that (3) is violated while (2) and (4) hold.*

2.3. Step-wise procedures

First define a critical value sequence as any nondecreasing sequence $\tau = (\tau_k)_{1\leq k\leq m} \in [0, 1]^m$ (with $\tau_0 = 0$ by convention).

The *step-up procedure* of critical value sequence τ, denoted by $\mathbf{SU}(\tau)$, rejects the i-th hypothesis if $p_i \leq \tau_{\widehat{k}}$, with $\widehat{k} = \max\{k \in \{0, 1, ..., m\} : p_{(k)} \leq \tau_k\}$, where $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ denote the ordered p-values (with the convention $p_{(0)} = 0$).

The *step-down procedure* of critical value sequence τ, denoted by $\mathbf{SD}(\tau)$, rejects the i-th hypothesis if $p_i \leq \tau_{\widetilde{k}}$, with $\widetilde{k} = \max\{k \in \{0, 1, ..., m\} : \forall k' \leq k,\ p_{(k')} \leq \tau_{k'}\}$. It is straightforward to check that, for the same set of critical values, the step-up version always rejects more hypotheses than the step-down version. More comments and illustrations on step-wise procedures can be found in Blanchard et al. (2014) and Dickhaus (2014), among others.

2.4. False discovery rate

We measure the quantity of false positives of a step-up (resp. step-down) procedure by using the false discovery rate (FDR), introduced and popularised by Benjamini and Hochberg (1995), which is defined as the averaged proportion of errors among the rejected hypotheses. More formally, for some procedure R

rejecting the i-th hypothesis if $p_i \leq \hat{t}(X)$ (for some threshold $\hat{t}(X)$), we let

$$\mathrm{FDR}(R,P) = \mathbf{E}_{X\sim P}\left[\frac{\sum_{i\in\mathcal{H}_0(P)}\mathbf{1}\{p_i \leq \hat{t}(X)\}}{1\vee\sum_{i=1}^{m}\mathbf{1}\{p_i \leq \hat{t}(X)\}}\right],\ P\in\mathcal{P}. \tag{5}$$

The main contribution of this work is to propose procedures that control the FDR at a prescribed level α and that incorporate the knowledge of the F_i's in a way that increases the number of discoveries.

3. Procedures

In this section we briefly review some existing methods for FDR control and introduce our new procedures.

3.1. Existing methods

We use the following methods as starting points for constructing new procedures.

- [BH]: the seminal procedure proposed in Benjamini and Hochberg (1995), corresponding to the step-up procedure $\mathbf{SU}(\tau)$, with critical values $\tau_k = \alpha k/m$, $1\leq k\leq m$;
- [BR-λ], $\lambda\in(0,1)$: an adaptive version of the BH procedure that was proposed in Blanchard and Roquain (2009), corresponding to the step-up procedure $\mathbf{SU}(\tau)$, with critical values

$$\tau_k = \left((1-\lambda)\frac{\alpha k}{m-k+1}\right)\wedge\lambda,\ 1\leq k\leq m; \tag{6}$$

- [GBS]: an adaptive version of the BH procedure that has been proposed in Gavrilov et al. (2009), corresponding to the step-down procedure $\mathbf{SD}(\tau)$, with critical values

$$\tau_k = \frac{\alpha k}{m-(1-\alpha)k+1},\ 1\leq k\leq m; \tag{7}$$

- [Heyse]: the step-up procedure $\mathbf{SU}(\tau)$ using critical values given by

$$\tau_k = \max\{t\in\mathcal{A}\,:\,\overline{F}(t)\leq\alpha k/m\},\ 1\leq k\leq m; \tag{8}$$

where $\overline{F}$ is defined by (1). This procedure was proposed in Heyse (2011).

The rationale behind the critical values of [BR-λ] and [GBS] is that they are intended to mimic the oracle critical values $\tau_k = \alpha k/m_0(P)$, $1\leq k\leq m$, which are less conservative than those of [BH] when $m_0(P)/m$ is not close to 1, see, e.g., Benjamini et al. (2006); Blanchard and Roquain (2009) for more details on this issue. Also, among adaptive procedures, [GBS] satisfies a kind of optimality as a finite sample version of the asymptotically optimal rejection curve, see Finner et al. (2009).

Let us now comment on [Heyse]. First, in the continuous setting where (2) holds, $\overline{F}(t) \leq t$, $t \in [0,1]$, and thus the critical values given by (8) satisfy $\tau_k \geq \alpha k/m$, $1 \leq k \leq m$, which means that [Heyse] rejects at least as many hypotheses as [BH]. When (3) additionally holds, we have $\overline{F}(t) = t$, $t \in [0,1]$, and the two critical value sequences are the same. Second, in the discrete setting where (2) holds, $\mathcal{A}$ is finite and τ_k is not necessarily greater than $\alpha k/m$ anymore. However, [Heyse] is also less conservative (or equal) than [BH] in the latter case, as stated in the following result (proved in Appendix B for completeness).

Lemma 1. *Consider the model of Section 2.1 assuming (2), both in the continuous and discrete setting described in Section 2.2. Then the set of nulls rejected by [Heyse] is larger than the one of [BH] (almost surely). Furthermore, under (4), these two rejection sets are equal (almost surely) if $F_i = F_j$ for all $i \neq j$.*

The equality case of Lemma 1 was provided in Proposition 2.3 of Heller and Gur (2011). It can be seen as a limitation of Heyse procedure in the homogeneous case. In the heterogeneous case, however, $\overline{F}(t)$ is smaller than t (see Figure 1) and [Heyse] can substantially improve [BH] (see Figure 2).

While [Heyse] incorporates the knowledge of the F_i's in a natural way (see also Remark 1 below), it is not correctly calibrated for a rigorous FDR control (see Appendix B.1). We propose suitable modifications of [Heyse] in the next sections.

Remark 1 (Empirical Bayes point of view on the Heyse procedure)**.**
We claim that [Heyse] corresponds to a suitable empirical Bayes procedure. To see this, consider the "binomial example" of Section 2.2, but assume now that the counts $n_1, \dots, n_m$ are observed from a sample $N_1, \dots, N_m$ i.i.d. of an a priori distribution ν. Unconditionally, the p-values p_i, $i \in \mathcal{H}_0$, are thus i.i.d. with c.d.f. $\bar{F}_0 = \sum_{n \geq 0} \nu(\{n\}) F_{0,n}$, where $F_{0,n}$ is the c.d.f. jumping at each $x_{k,n} = 2^{-n} \sum_{j=0}^{k-1} \binom{n}{j}$ with $F_{0,n}(x_{k,n}) = x_{k,n}$, $1 \leq k \leq n+1$. This suggests to normalise the p-values p_i as $\bar{F}_0(p_i)$ which leads to the step-up procedure with critical values $\tau_k = \max\{t : \bar{F}_0(t) \leq \alpha k/m\}$. Following an empirical Bayes approach, the prior ν can be estimated by $\hat{\nu}(\{n\}) = m^{-1} \sum_{i=1}^{m} \mathbf{1}_{\{N_i = n\}}$, which gives rise to the estimator of $\bar{F}_0$ defined by $\hat{\bar{F}}_0 = \sum_{n \geq 0} \hat{\nu}(\{n\}) F_{0,n} = m^{-1} \sum_{i=1}^{m} F_{0,N_i}$, which is equal to $\overline{F}$ given by (1). Hence, the corresponding (empirical Bayes) step-up procedure reduces to [Heyse].

3.2. Two new methods

We now present two procedures that aim at correcting [Heyse] :

- [HSU] (heterogeneous step-up) : the step-up procedure $\mathbf{SU}(\tau)$ using the critical values defined in the following way:

$$\tau_m = \max\left\{t \in \mathcal{A} \,:\, \frac{1}{m}\sum_{i=1}^{m}\frac{F_i(t)}{1-F_i(t)} \le \alpha\right\} \tag{9}$$

$$\tau_k = \max\left\{t \in \mathcal{A} \,:\, t \le \tau_m,\ \frac{1}{m}\sum_{i=1}^{m}\frac{F_i(t)}{1-F_i(\tau_m)} \le \alpha k/m\right\},\ 1 \le k \le m-1. \tag{10}$$

- [HSD] (heterogeneous step-down) : the step-down procedure $\mathbf{SD}(\tau)$ using the critical values defined in the following way :

$$\tau_k = \max\left\{t \in \mathcal{A} \,:\, \frac{1}{m}\sum_{i=1}^{m}\frac{F_i(t)}{1-F_i(t)} \le \alpha k/m\right\},\ 1 \le k \le m. \tag{11}$$

[HSU] can be seen as a correction of [Heyse]: the correction term in the critical values (10) lies in the additional denominator $1-F_i(\tau_m)$. A consequence is that [HSU] can be more conservative than [BH]. However, the magnitude of this phenomenon is always small, as the next lemma shows (proved in Appendix B for completeness).

Lemma 2. *Under the conditions of Lemma 1, the set of nulls rejected by [HSU] contains the one of [BH] taken at level* $\alpha/(1+\alpha)$ *(almost surely).*

For [HSD], the following result can be established.

Lemma 3. *Under the conditions of Lemma 1, the set of nulls rejected by [HSD] contains the one of the step-down procedure with critical values* $(\alpha k/m)/(1+\alpha k/m)$, $1 \le k \le m$ *(almost surely).*

From (10) and (11) it is clear that the critical values of [HSD] are always at least as large as those for [HSU]. However, since the step-up direction is more powerful than the step-down direction (see Section 2.3) neither of the two generally dominates the other one.

Remark 2. *We may ask whether we can construct a uniform improvement of [BH] that incorporates the* F_i*'s. There is indeed such a procedure, see procedure [RBH] in Appendix A.1 for more details. However, the improvement brought by the* F_i*'s information is less substantial than for [HSU], so we have chosen to omit [RBH] from the main stream of the paper.*

3.3. Adaptive versions

In this section, we define adaptive versions of [HSU] and [HSD] in the following way:

- [AHSU] (one-stage adaptive heterogeneous step-up): the step-up procedure $\mathbf{SU}(\tau)$ using the critical values defined in the following way: τ_m as

in (9) and for $1 \leq k \leq m-1$,

$$\tau_k = \max\left\{t \in \mathcal{A} \,:\, t \leq \tau_m,\ \left(\frac{F(t)}{1-F(\tau_m)}\right)_{(1)} + \cdots + \left(\frac{F(t)}{1-F(\tau_m)}\right)_{(m-k+1)} \leq \alpha k\right\}, \tag{12}$$

where each $\left(\frac{F(t)}{1-F(\tau_m)}\right)_{(j)}$ denotes the j-th largest element of the range of values $\left\{\frac{F_i(t)}{1-F_i(\tau_m)}, 1 \leq i \leq m\right\}$.

- [AHSD] (one-stage adaptive heterogeneous step-down): the step-down procedure $\mathbf{SD}(\tau)$ using the critical values defined in the following way: for $1 \leq k \leq m$,

$$\tau_k = \max\left\{t \in \mathcal{A} \,:\, \left(\frac{F(t)}{1-F(t)}\right)_{(1)} + \cdots + \left(\frac{F(t)}{1-F(t)}\right)_{(m-k+1)} \leq \alpha k\right\}, \tag{13}$$

where each $\left(\frac{F(t)}{1-F(t)}\right)_{(j)}$ denotes the j-th largest elements of the range of values $\left\{\frac{F_i(t)}{1-F_i(t)}, 1 \leq i \leq m\right\}$.

Note that the critical values of [AHSU] and [AHSD] are clearly larger than or equal to those of their non-adaptive counterparts [HSU] and [HSD], respectively. This means that the adaptive versions are always less conservative. The following result establishes a connection of the adaptive procedures to the [BR-λ] and [GBS] procedures (proved in Appendix B for completeness).

Lemma 4. *Under the conditions of Lemma 1, the following holds:*

(i) the set of nulls rejected by [AHSU] contains the one of [BR-λ] (a.s.) for λ equals to (9);
(ii) the set of nulls rejected by [AHSD] contains the one of [GBS] (a.s.);

The above lemma ensures that the user can incorporate the knowledge of the F_i's in adaptive procedures with a "no loss" guarantee with respect to [BR] and [GBS].

Remark 3. *We may ask whether we can build a procedure that is a uniform improvement of [BR-λ], for any fixed value of $\lambda \in (0,1)$. We propose a solution in Appendix A.2, called [HBR-λ]. It does not improve uniformly [HSU], but is an interesting variant of [AHSU].*

4. New FDR bounds for heterogeneous nulls

In this section, we present new FDR bounds which are the main mathematical contributions of this paper and that are of independent interest. They generalise

some classical bounds from super-uniform null distributions to arbitrary heterogeneous (not necessarily discrete) null distributions, and immediately yield FDR control of our new procedures.

4.1. Results

The following result holds. It only assumes independence between the p-values and *not* super-uniformity of the null distributions.

Theorem 1. *Consider any family $\mathcal{F} = \{F_i, 1 \le i \le m\}$ as defined in Section 2.1 and assume* (Indep)*. Consider any critical values τ_k, $1 \le k \le m$ such that $\forall i \in \{1, \dots, m\}$, $F_i(\tau_m) < 1$. Then, for all $P \in \mathcal{P}$, we have*

$$\begin{aligned}&\mathrm{FDR}(\mathbf{SU}(\tau), P)\\ &\le \min\left(\sum_{i=1}^{m} \max_{1\le k\le m} \frac{F_i(\tau_k)}{k}, \max_{1\le k\le m} \max_{\substack{A\subset\{1,\dots,m\}\\ |A|=m-k+1}} \left(\frac{1}{k}\sum_{i\in A} \frac{F_i(\tau_k)}{1-F_i(\tau_m)}\right)\right); \quad (14)\\ &\mathrm{FDR}(\mathbf{SD}(\tau), P)\\ &\le \min\left(\sum_{i=1}^{m} \max_{1\le k\le m} \frac{F_i(\tau_k)}{k}, \max_{1\le k\le m} \max_{\substack{A\subset\{1,\dots,m\}\\ |A|=m-k+1}} \left(\frac{1}{k}\sum_{i\in A} \frac{F_i(\tau_k)}{1-F_i(\tau_k)}\right)\right). \quad (15)\end{aligned}$$

The proof of Theorem 1 is deferred to Section 8. It combines several techniques: the first tool is an expression of the FDR introduced by Ferreira (2007) (step-up case) and Roquain and Villers (2011) (step-down case). A second idea comes from the work Blanchard and Roquain (2009) (step-up case) and Gavrilov et al. (2009) (step-down case), which introduced a new term (here, the denominator $(1 - F_i(\cdot))$) to make the proof work. Finally, another inspiration is the study of Roquain and van de Wiel (2009) and Döhler (2016) that allowed to deal with heterogeneous FDR thresholding. Let us underline that the obtained proof is especially concise, which means that these different techniques fit together perfectly well, which is perhaps surprising at first glance, see Section 8.

Next, let us note that taking the maximum over the subset A in (14) and (15) allows us to adapt to the unknown number of true null hypotheses: loosely, if $k-1$ is the number of rejections, A corresponds to the acceptation set (hence of cardinality $m-k+1$), which "estimates" $\mathcal{H}_0$ and thus the sums in (14) and (15) are indexed by a set "close" to the unknown set $\mathcal{H}_0$. Taking the maximum then corresponds to the least favorable possible $\mathcal{H}_0$.

Finally, let us underline again that the above bounds do not use the super-uniformity of the F_i's which makes them quite general and flexible tools. Several examples are given below.

4.2. Application to adaptiveness and weighting

Let us now give some intuition behind these bounds and illustrate their generality by showing how they encompasse previous work in the literature. First, assuming the super-uniformity $F_i(t) \leq t$ for all t and i, these bounds entail

$$\mathrm{FDR}(\mathbf{SU}(\tau), P) \leq m \max_{1\leq k\leq m} \{\tau_k/k\}; \tag{16}$$

$$\mathrm{FDR}(\mathbf{SU}(\tau), P) \leq \max_{1\leq k\leq m} \frac{m-k+1}{1-\tau_m}\frac{\tau_k}{k}; \tag{17}$$

$$\mathrm{FDR}(\mathbf{SD}(\tau), P) \leq \max_{1\leq k\leq m} \frac{m-k+1}{1-\tau_k}\frac{\tau_k}{k}, \tag{18}$$

which immediately implies that [BH], [BR-λ] (with $\tau_m = \lambda$) and [GBS] all control the FDR at level α. To this respect, bounds (16), (17) and (18) encompass Theorem 1 of Benjamini and Hochberg (1995), Theorem 9 of Blanchard and Roquain (2009) and Theorem 1.1 of Gavrilov et al. (2009), respectively.

Second, by removing the adaptative part of the bounds, that is, by replacing A by $\{1,\dots,m\}$, we obtain the simpler but more conservative bounds

$$\mathrm{FDR}(\mathbf{SU}(\tau), P) \leq \max_{1\leq k\leq m} \left(\frac{1}{k}\sum_{i=1}^{m} \frac{F_i(\tau_k)}{1-F_i(\tau_m)}\right); \tag{19}$$

$$\mathrm{FDR}(\mathbf{SD}(\tau), P) \leq \max_{1\leq k\leq m} \left(\frac{1}{k}\sum_{i=1}^{m} \frac{F_i(\tau_k)}{1-F_i(\tau_k)}\right). \tag{20}$$

Here, we show how these bounds can be used to recover some of the finite sample FDR controlling results of Roquain and van de Wiel (2009) for p-value weighting procedures. Assume that the p-values p_i, $1 \leq i \leq m$, have uniform marginals under the null, that is, satisfy (3) and consider any family of c.d.f. $(\Delta_i)_{1\leq i\leq m}$, with the additional property $m^{-1}\sum_{i=1}^{m}\Delta_i(x) = x$, for $x \in [0,\alpha]$. This family can be considered as "weighting" the p-values. It is a free parameter that adds an extra flexibility which can be useful in different contexts, see, e.g., Ignatiadis et al. (2016), Durand (2017). An important point is then to make sure that this weighting maintains the FDR control. For this, let us first modify the family $(\Delta_i)_{1\leq i\leq m}$ as follows:

$$\tilde{\Delta}_i(x) = \frac{\Delta_i(x)}{1+\Delta_i(\alpha)}, \quad \text{so that} \quad \frac{\tilde{\Delta}_i(x)}{1-\tilde{\Delta}_i(\alpha)} = \Delta_i(x), \ \ x \in [0,1), \ \ 1 \leq i \leq m,$$

with the convention $\tilde{\Delta}_i(1) = 1$ (to make $\tilde{\Delta}_i$ meet the properties of a c.d.f.). Then we can consider the BH procedure using the transformed p-values $\tilde{p}_i = \tilde{\Delta}_i^{-1}(p_i)$, $1 \leq i \leq m$, which can be interpreted as a "weighted BH procedure", in the sense that each p-value p_i has an importance which is increased or diminished in the procedure according to the value of $\tilde{\Delta}_i^{-1}$ at p_i. Since each $\tilde{p}_i$ has for null c.d.f.

$\tilde{\Delta}_i$, our bound (19) yields

$$\mathrm{FDR} \leq \max_{1\leq k\leq m}\left(\frac{1}{k}\sum_{i=1}^{m}\frac{\tilde{\Delta}_i\left(\alpha k/m\right)}{1-\tilde{\Delta}_i\left(\alpha\right)}\right) = \max_{1\leq k\leq m}\left(\frac{1}{k}\sum_{i=1}^{m}\Delta_i\left(\alpha k/m\right)\right) = \alpha,$$

which recovers the results of Theorem 4.1 in Roquain and van de Wiel (2009) (step-up part). The step-down part can be recovered from (20) in a similar way.

4.3. Application to the new procedures

To make a connection between Theorem 1 and our new procedures, especially [AHSU] and [AHSD] (see Section 3.3), observe that the following relations hold true:

$$\max_{\substack{A\subset\{1,\dots,m\}\\|A|=m-k+1}}\left(\sum_{i\in A}\frac{F_i\left(\tau_k\right)}{1-F_i\left(\tau_m\right)}\right) = \left(\frac{F\left(\tau_k\right)}{1-F\left(\tau_m\right)}\right)_{(1)} + \cdots + \left(\frac{F\left(\tau_k\right)}{1-F\left(\tau_m\right)}\right)_{(m-k+1)};$$

$$\max_{\substack{A\subset\{1,\dots,m\}\\|A|=m-k+1}}\left(\sum_{i\in A}\frac{F_i\left(\tau_k\right)}{1-F_i\left(\tau_k\right)}\right) = \left(\frac{F\left(\tau_k\right)}{1-F\left(\tau_k\right)}\right)_{(1)} + \cdots + \left(\frac{F\left(\tau_k\right)}{1-F\left(\tau_k\right)}\right)_{(m-k+1)}.$$

Therefore, Theorem 1 implies that our new procedures enjoy the desired FDR controlling property.

Corollary 1. *In the setting of Theorem 1, the procedures [HSU], [HSD], [AHSU], [AHSD] all control the FDR at level* α.

Now let us focus on the discrete case. In that situation, recall that the individual p-values cannot be transformed (without randomisation) to be uniform under the null. Rather, our Heyse-type procedures "average" the heterogeneous nulls. As a consequence, if some of the F_i's are really small, they will not contribute much to the average, offering some additional room for the other F_j's.

Finally, let us underline that our bounds can be useful for other discrete-type procedures. As a case in point, consider mid p-values which were introduced by Lancaster (1961) and are sometimes used for analysing discrete data (see, e.g., Karp et al., 2016). These p-values are no longer super-uniform under the null hypotheses, however our theorem can accommodate such distributions in a natural way to still yield valid FDR controlling procedures.

5. Empirical data

To illustrate the performance of FDR-controlling procedures for discrete data, we analyse two classical data sets. In what follows, our main goal is to compare the performance of the new procedures [HSU] and [AHSU] to the classical [BH] and [Storey] and also to [Heyse]. The procedure [Storey] was proposed in Storey

et al. (2004), and corresponds to the step-up procedure $\mathbf{SU}(\tau)$, with critical values $\tau_k = \alpha k/\widehat{m}_0$, $1 \leq k \leq m$, where

$$\widehat{m}_0 = \widehat{m}_0(\lambda) = \frac{1 + \sum_{i=1}^{m} \mathbf{1}\{p_i > \lambda\}}{1 - \lambda}$$

is an estimate of the number m_0 of true null hypotheses among the m hypotheses. We use the standard value $\lambda = \frac{1}{2}$. All analyses were performed using the R language for statistical computing (R Core Team, 2016).

5.1. Pharmacovigilance data

This data set is derived from a database for reporting, investigating and monitoring adverse drug reactions due to the Medicines and Healthcare products Regulatory Agency in the United Kingdom. It contains the number of reported cases of amnesia as well as the total number of adverse events reported for each of the $m = 2446$ drugs in the database. For more details we refer to Heller and Gur (2011) and to the accompanying R-package 'discreteMTP' (Heller et al., 2012), which also contains the data. Heller and Gur (2011) investigate the association between reports of amnesia and suspected drugs by performing for each drug a Fisher's exact test (one-sided) for testing association between the drug and amnesia while adjusting for multiplicity by using several (discrete) FDR procedures.

5.2. Next generation sequencing data

We also revisit the next generation sequencing (NGS) count data analysed by Chen and Doerge (2015b), to which we also refer for more details. More specifically, we reanalyse the methylation data set for cytosines of Arabidopsis in Lister et al. (2008) which is part of the R-package 'fdrDiscreteNull' (Chen and Doerge, 2015a). This data set contains the counts for a biological entity under two different biological conditions or treatments. Following Chen and Doerge (2015b), $m = 7421$ genes whose treatment-wise total counts are positive but row-total counts are no greater than 100 are analysed using the exact binomial test, see Chen and Doerge (2015b).

5.3. Results

Table 1 summarises the number of discoveries for the pharmacovigilance and NGS data when using the respective FDR procedures at level $\alpha = 0.05$. Compared to the classical [BH] procedure, the discrete procedures are able to detect three additional candidates linking amnesia and drugs in the pharmacovigilance data. This data set seems to contain very few signals so there is no benefit in using adaptive procedures, in fact the (finite sample) [Storey] procedure performs worse than the [BH] procedure. Note also that our new procedures – while being correctly calibrated – still reject the same number of hypotheses as [Heyse].

TABLE 1
Number of rejections (discoveries) for the pharmacovigilance and Arabidopsis methylation data.

Data set	[BH]	[HSU]	[Heyse]	[Storey]	[AHSU]
Pharmacovigilance	24	27	27	22	27
Arabidopsis methylation	2097	2358	2379	2395	2446

In contrast, the Arabidopsis data seems to contain a large portion of signals so that in particular the [Storey] procedure performs much better than [BH]. The [HSU] and [Heyse] procedures also outperform [BH], while the [Storey] procedure is dominated by the [AHSU] procedure.

Figure 2 illustrates graphically the data and the critical constants of the involved multiple testing procedures. In particular, the benefit of taking dis-

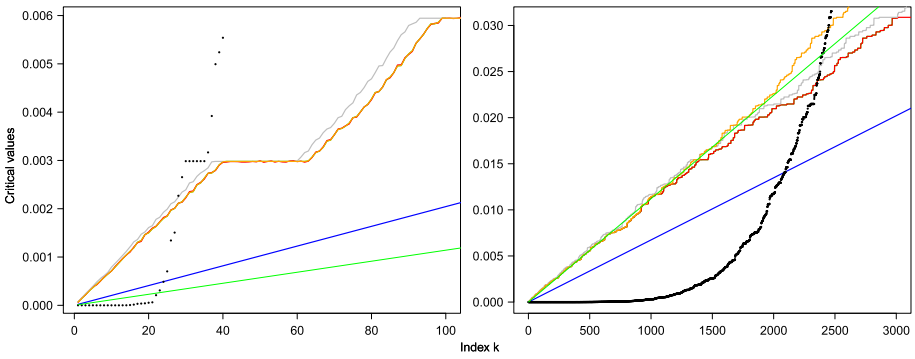


FIG 2. *Critical constants and sorted p-values (represented by black dots) for the pharmacovigilance (left panel) and Arabidopsis methylation data (right panel). The [BH], [HSU], [Heyse], [Storey] and [AHSU] critical constants are represented respectively by blue, red, grey, green, and orange solid lines.*

creteness into account becomes more apparent: for the pharmacovigilance data, the discrete critical values are considerably (by a factor of $2.5-3.5$) larger than their respective classical counterparts. This leads to more powerful procedures. For the NGS data, we can observe quite clearly that the [HSU] critical constants are dominated by the [AHSU] constants, as explained in Section 3. This leads to roughly 100 additional rejections. Again, the discrete critical values are considerably larger than their respective classical counterparts. In Section 3.2, we mentioned that the correction factor $1 - F_i(\tau_m)$, introduced for guaranteeing FDR control of [HSU], may lead to a procedure which is more conservative than [BH]. However, Figure 2 shows that – at least for the data sets considered here – this risk is by far compensated by the benefit of taking discreteness adequately into account.

6. Simulation study

We now investigate the power of the procedures from the previous section in a simulation study similar to those described in Gilbert (2005), Heller and Gur (2011) and Döhler (2016). Again, we focus on comparing the performance of the new discrete procedures to [BH], [Storey] and [Heyse].

6.1. Simulated scenarios

We simulate a two-sample problem in which a vector of m independent binary responses ("adverse events") is observed for each subject in two groups, where each group consists of $N = 25$ subjects. Then, the goal is to simultaneously test the m null hypotheses H_{0i} : "$p_{1i} = p_{2i}$", $i = 1, \ldots, m$, where p_{1i} and p_{2i} are the success probabilities for the ith binary response in group 1 and 2, respectively. Before we describe the simulation framework in more detail, we explain how this set-up leads to discrete and heterogeneous p-value distributions. Suppose we have simulated two vectors of dimension m where each component represents a count in $\{0, \ldots, 25\}$. This data can be represented by m contingency tables. Now each hypothesis is tested using Fisher's exact test (two-sided) for each contingency table, which is performed by conditioning on the (simulated) pair of marginal counts. Thus, we can determine for every contingency table i the discrete distribution function F_i of the p-values for Fisher's exact test under the null hypothesis. For differing (simulated) contingency tables, these induced distributions will generally be heterogeneous and our inference is conditionally on the marginal counts.

We take $m = 800, 2000$ where $m = m_1 + m_2 + m_3$ and data are generated so that the response is $Bernoulli(0.01)$ at m_1 positions for both groups, $Bernoulli(0.10)$ at m_2 positions for both groups and $Bernoulli(0.10)$ at m_3 positions for group 1 and $Bernoulli(q)$ at m_3 positions for group 2 where $q = 0.15, 0.25, 0.4$ represents weak, moderate and strong effects respectively. The null hypothesis is true for the m_1 and m_2 positions while the alternative hypothesis is true for the m_3 positions. We also take different configurations for the proportion of false null hypotheses, m_3 is set to be 10%, 30% and 80% of the value of m, which represents small, intermediate and large proportion of effects (the proportion of true nulls π_0 is 0.9, 0.7, 0.2, respectively). Then, m_1 is set to be 20%, 50% and 80% of the number of true nulls (that is, $m - m_3$) and m_2 is taken accordingly as $m - m_1 - m_3$.

For each of the 54 possible parameter configurations specified by m, m_3, m_1 and q, 10000 Monte Carlo trials are performed, that is, 10000 data sets are generated and for each data set, an unadjusted two-sided p-value from Fisher's exact test is computed for each of the m positions, and the multiple testing procedures mentioned above are applied at level $\alpha = 0.05$. The power of each procedure was estimated as the fraction of the m_3 false null hypotheses that were rejected, averaged over the 10000 simulations. Note that while our procedures are designed to control the FDR conditionally on the marginal counts,

our power results are presented in an unconditional way for the sake of simplicity. For random number generation the R-function *rbinom* was used. The two-sided p-values from Fisher's exact test were computed using the R-function *fisher.test*.

6.2. Results

We have computed the (average) power and FDR of the five procedures under investigation in all scenarios (see Tables 3 and 4 in Appendix E for the full display). For weak and moderate effects, i.e. $q = 0.15$ and $q = 0.25$, none of the procedure possesses relevant power. For strong effects, the results are summarised in Figure 3. (Since the power of the discrete procedures is slightly increasing in m_1 for fixed m_3 and q, we present – in order to avoid over-optimism – the configuration with smallest m_1).

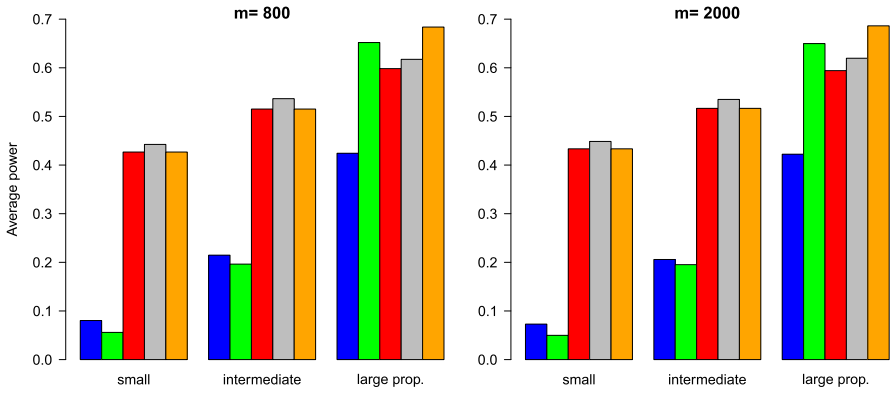


FIG 3. *Average power for the [BH], [Storey], [HSU], [Heyse] and [AHSU] procedures in the simulation study. The coloring is the same as in Figure 2.*

The results are consistent with the findings of the previous section: the new discrete procedures are considerably more powerful than the [BH] procedure. When the proportion of alternatives is large, the [Storey] procedure provides large gains over [BH] but is still dominated by the discrete adaptive procedure [AHSU].

7. Conclusion and discussion

In this paper, we provided new bounds for the FDR of step-up and step-down procedures that use heterogeneous test statistics. This made it possible to define a new class of multiple testing procedures that provably control the FDR while incorporating the discreteness and heterogeneity of the tests statistics in a convenient way. We have shown that our approach can be seen as cor-

recting and improving the approach of Heyse (2011): while it ensures a theoretical control, it can also make more rejections when the signal is strong enough.

Our new procedures are easily interpreted since they involve neither randomisation nor any additional choice of tuning parameters (in Appendix C we present a comparison with a randomised p-value approach). Furthermore, an R-package implementing them is currently being developed, which will make these methods available for the practitioner.

Additionally, our methodology can deal with other null distributions F_i that arise in the context of discrete testing: as a case in point, consider mid p-values which were introduced by Lancaster (1961) and are sometimes used for analysing discrete data (see, e.g., Karp et al. (2016)). These p-values are no longer super-uniform under the null hypotheses, however our methods can accommodate such distributions in a natural way to still yield valid FDR controlling procedures.

Finally, this paper opens several directions for future research, especially by trying to extend our arguments to other frameworks. For instance, an important point is to relax the independence requirement. To this respect, we believe that our procedures will inherit the behavior of BH procedure: while the FDR control is likely to be maintained under "realistic" dependence, formally proving such a result is probably a challenging problem. Another challenge is to develop mathematically valid plug-in procedures for discrete data. A first step in this direction is sketched in Appendix D.

8. Proof of Theorem 1

8.1. Lemmas for step-down and step-up procedures

Let us introduce the following modifications of $\mathbf{SU}(\tau)$:

- $\mathbf{SU}^{\sharp}(\tau) = \mathbf{SU}(\tau^{\sharp})$ the step-up with m critical values defined by $(\tau_1^{\sharp}, \ldots, \tau_m^{\sharp}) = (\tau_2, \ldots, \tau_m, \tau_m)$;
- for some given index $i \in \{1, \ldots, m\}$, $\mathbf{SU}^{\sharp,-i}(\tau) = \mathbf{SU}(\tau^{\sharp,-i})$ the step-up with $m-1$ critical values defined by $(\tau_1^{\sharp,-i}, \ldots, \tau_{m-1}^{\sharp,-i}) = (\tau_2, \ldots, \tau_m)$ and restricted to the p-values of the set $\{p_j, j \neq i\}$.

The following lemma holds (variation of a well known lemma, see, e.g., Ferreira and Zwinderman, 2006) and is proved in Appendix B for completeness.

Lemma 5. *For all $i \in \{1, \ldots, m\}$, the following assertions are equivalent: (i) $p_i \leq \tau_{\widehat{k}}$; (ii) $p_i \leq \tau_{\hat{k}^{\sharp,-i}+1}$; (iii) $\hat{k}^{\sharp,-i} + 1 = \widehat{k}$, where $\hat{k}^{\sharp,-i}$ denotes the number of rejected hypotheses of the procedure $\mathbf{SU}^{\sharp,-i}(\tau)$. Moreover, we have $\{p_i > \tau_m\} \subset \{\hat{k}^{\sharp} = \hat{k}^{\sharp,-i}\}$, where $\hat{k}^{\sharp}$ denotes the number of rejected hypotheses of the procedure $\mathbf{SU}^{\sharp}(\tau)$.*

Let us introduce the following modifications of $\mathbf{SD}(\tau)$:

- for some given index $i \in \{1,\dots,m\}$, $\mathbf{SD}^{-i}(\tau) = \mathbf{SD}(\tau^{-i})$ the step-down procedure with $m-1$ critical values defined by $(\tau_1^{-i},\dots,\tau_{m-1}^{-i}) = (\tau_1,\dots,\tau_{m-1})$ and restricted to the p-values of the set $\{p_j, j \neq i\}$.
- for some given index $i \in \{1,\dots,m\}$, $\mathbf{SD}^{\sharp,-i}(\tau) = \mathbf{SD}(\tau^{\sharp,-i})$ the step-down procedure with the $m-1$ critical values $(\tau_1^{\sharp,-i},\dots,\tau_{m-1}^{\sharp,-i}) = (\tau_2,\dots,\tau_m)$ and restricted to the p-values of the set $\{p_j, j \neq i\}$.

The following lemma holds (variation of Gavrilov et al., 2009; Roquain and Villers, 2011) and is proved in Appendix B for completeness:

Lemma 6. *For all $i \in \{1,\dots,m\}$, the following assertions are equivalent: (i) $p_i \leq \tau_{\tilde{k}}$; (ii) $p_i \leq \tau_{\tilde{k}+1}$; (iii) $p_i \leq \tau_{\tilde{k}^{-i}+1}$; (iv) $\tilde{k}^{\sharp,-i}+1 = \tilde{k}$, where $\tilde{k}^{-i}$ is the number of rejections of $\mathbf{SD}^{-i}(\tau)$ and $\tilde{k}^{\sharp,-i}$ is the number of rejections of $\mathbf{SD}^{\sharp,-i}(\tau)$. Moreover, we have $\{p_i > \tau_{\tilde{k}^{-i}+1}\} \subset \{\widetilde{k} = \tilde{k}^{-i}\}$.*

8.2. Proof of Theorem 1, step-up part

By using Lemma 5 (ii) and (iii), we obtain

$$\mathrm{FDR}(\mathbf{SU}(\tau)) = \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\hat{k}}\}}{\widehat{k}}\right) = \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\hat{k}^{\sharp,-i}+1}\}}{\hat{k}^{\sharp,-i}+1}\right). \quad (21)$$

because $p_i \leq \tau_{\hat{k}}$ is equivalent to $p_i \leq \tau_{\hat{k}^{\sharp,-i}+1}$, and both imply $\hat{k}^{\sharp,-i}+1 = \widehat{k}$. Now using independence between $\hat{k}^{\sharp,-i}$ and p_i, we obtain

$$\begin{aligned}
\sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\hat{k}^{\sharp,-i}+1}\}}{\hat{k}^{\sharp,-i}+1}\right) &= \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\hat{k}^{\sharp,-i}+1}\}}{\hat{k}^{\sharp,-i}+1} \,\middle|\, \hat{k}^{\sharp,-i}\right)\right) \\
&= \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{\mathbf{P}\left(p_i \leq \tau_{\hat{k}^{\sharp,-i}+1} \mid \hat{k}^{\sharp,-i}\right)}{\hat{k}^{\sharp,-i}+1}\right) \\
&\leq \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp,-i}+1}\right)}{\hat{k}^{\sharp,-i}+1}\right),
\end{aligned}$$

because for any $i \in \mathcal{H}_0$, and t, we have $\mathbf{P}(p_i \leq t) \leq F_i(t)$. Now, on the one hand,

$$\sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp,-i}+1}\right)}{\hat{k}^{\sharp,-i}+1}\right) \leq \sum_{i=1}^{m} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp,-i}+1}\right)}{\hat{k}^{\sharp,-i}+1}\right) \leq \sum_{i=1}^{m} \max_{1\leq k\leq m} \frac{F_i(\tau_k)}{k}.$$

Next, on the other hand, by using again (Indep) and that for any $i \in \mathcal{H}_0$, and t, $1 - \mathbf{P}(p_i \leq t) \geq 1 - F_i(t)$,

$$\begin{aligned}\sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp,-i}+1}\right)}{\hat{k}^{\sharp,-i}+1}\right) &\leq \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp,-i}+1}\right)}{\hat{k}^{\sharp,-i}+1}\mathbf{E}\left(\frac{\mathbf{1}\{p_i > \tau_m\}}{1-F_i(\tau_m)} \,\middle|\, \hat{k}^{\sharp,-i}\right)\right) \\ &= \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp,-i}+1}\right)}{1-F_i(\tau_m)}\frac{\mathbf{1}\{p_i > \tau_m\}}{\hat{k}^{\sharp,-i}+1}\right) \\ &\leq \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\hat{k}^{\sharp}+1}\right)}{1-F_i(\tau_m)}\frac{\mathbf{1}\{p_i > \tau_m\}}{\hat{k}^{\sharp}+1}\mathbf{1}\{\hat{k}^{\sharp}+1 \leq m\}\right),\end{aligned}$$

where the latter inequality comes from the last assertion of Lemma 5. Now, since $\tau_{\hat{k}^{\sharp}+1} \leq \tau_m$, we have that the last display is smaller than or equal to

$$\begin{aligned}&\mathbf{E}\left(\sum_{i\in\mathcal{H}_0}\frac{F_i\left(\tau_{\hat{k}^{\sharp}+1}\right)}{1-F_i(\tau_m)}\frac{\mathbf{1}\{p_i > \tau_{\hat{k}^{\sharp}+1}\}}{\hat{k}^{\sharp}+1}\mathbf{1}\{\hat{k}^{\sharp}+1 \leq m\}\right) \\ &\leq \max_{0\leq k\leq m-1}\max_{\substack{A\subset\{1,\ldots,m\}\\|A|=m-k}}\sum_{i\in A\cap\mathcal{H}_0}\frac{F_i(\tau_{k+1})}{1-F_i(\tau_m)}\frac{1}{k+1}, \qquad (22)\end{aligned}$$

by taking the maximum over all the possible realisations of the set $A = \{1 \leq i \leq m : p_i > \tau_{\hat{k}^{\sharp}+1}\} = \{1 \leq i \leq m : p_i > \tau^{\sharp}_{\hat{k}^{\sharp}}\}$ which is the index set corresponding to the non-rejected null hypotheses of $\mathbf{SU}(\tau^{\sharp})$ (the latter being by definition of cardinality $m - \hat{k}^{\sharp}$). This concludes the proof.

8.3. *Proof of Theorem 1, step-down part*

It is similar to the step-up case, with some subtle changes:

$$\mathrm{FDR}(\mathbf{SD}(\tau)) = \sum_{i\in\mathcal{H}_0}\mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\tilde{k}}\}}{\tilde{k}}\right) = \sum_{i\in\mathcal{H}_0}\mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\tilde{k}^{-i}+1}\}}{\tilde{k}^{\sharp,-i}+1}\right),$$

because $p_i \leq \tau_{\tilde{k}}$ is equivalent to $p_i \leq \tau_{\tilde{k}^{-i}+1}$, and both imply $\tilde{k}^{\sharp,-i}+1 = \tilde{k}$ (keep in mind that $\tilde{k}^{\sharp,-i}$ might be different from $\tilde{k}^{-i}$), by applying Lemma 6. Now using independence between $(\tilde{k}^{-i}, \tilde{k}^{\sharp,-i})$ and p_i, we obtain

$$\begin{aligned}\sum_{i\in\mathcal{H}_0}\mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\tilde{k}^{-i}+1}\}}{\tilde{k}^{\sharp,-i}+1}\right) &= \sum_{i\in\mathcal{H}_0}\mathbf{E}\left(\mathbf{E}\left(\frac{\mathbf{1}\{p_i \leq \tau_{\tilde{k}^{-i}+1}\}}{\tilde{k}^{\sharp,-i}+1} \,\middle|\, (\tilde{k}^{-i}, \tilde{k}^{\sharp,-i})\right)\right) \\ &= \sum_{i\in\mathcal{H}_0}\mathbf{E}\left(\frac{\mathbf{P}\left(p_i \leq \tau_{\tilde{k}^{-i}+1} \,\middle|\, (\tilde{k}^{-i}, \tilde{k}^{\sharp,-i})\right)}{\tilde{k}^{\sharp,-i}+1}\right)\end{aligned}$$

$$\leq \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\tilde{k}^{-i}+1}\right)}{\tilde{k}^{\sharp,-i}+1}\right)$$

$$\leq \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\tilde{k}^{-i}+1}\right)}{\tilde{k}^{-i}+1}\right),$$

because $\tilde{k}^{\sharp,-i}+1 \geq \tilde{k}^{-i}+1$ and because for $i \in \mathcal{H}_0$, and any t, we have $\mathbf{P}(p_i \leq t) \leq F_i(t)$. This gives the first part of the bound. Next, by using again (Indep), we obtain

$$\sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\tilde{k}^{-i}+1}\right)}{\tilde{k}^{-i}+1}\right) \leq \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\tilde{k}^{-i}+1}\right)}{\tilde{k}^{-i}+1}\mathbf{E}\left(\frac{\mathbf{1}\{p_i > \tau_{\tilde{k}^{-i}+1}\}}{1-F_i\left(\tau_{\tilde{k}^{-i}+1}\right)} \,\middle|\, (\tilde{k}^{-i}, \tilde{k}^{\sharp,-i})\right)\right)$$

$$= \sum_{i\in\mathcal{H}_0} \mathbf{E}\left(\frac{F_i\left(\tau_{\tilde{k}^{-i}+1}\right)}{1-F_i\left(\tau_{\tilde{k}^{-i}+1}\right)}\frac{\mathbf{1}\{p_i > \tau_{\tilde{k}^{-i}+1}\}}{\tilde{k}^{-i}+1}\right).$$

Now using the last assertion of Lemma 6, the last display is smaller than or equal to

$$\mathbf{E}\left(\sum_{i\in\mathcal{H}_0} \frac{F_i\left(\tau_{\tilde{k}+1}\right)}{1-F_i\left(\tau_{\tilde{k}+1}\right)}\frac{\mathbf{1}\{p_i > \tau_{\tilde{k}+1}\}}{\tilde{k}+1}\mathbf{1}\{\tilde{k}+1 \leq m\}\right)$$

$$\leq \mathbf{E}\left(\max_{0\leq k\leq m-1}\max_{\substack{A\subset\{1,\ldots,m\}\\|A|=m-k}}\sum_{i\in A\cap\mathcal{H}_0}\frac{F_i(\tau_{k+1})}{1-F_i(\tau_{k+1})}\frac{1}{k+1}\right),$$

because $\{1 \leq i \leq m : p_i > \tau_{\tilde{k}+1}\}$ is equal to $\{1 \leq i \leq m : p_i > \tau_{\tilde{k}}\}$, since both sets correspond to the set of non-rejected hypotheses of $\mathbf{SD}(\tau)$. Since $\mathbf{SD}(\tau)$ rejects exactly $\tilde{k}$ hypotheses, the proof is completed.

Appendix A: Additional procedures

A.1. A rescaled BH procedure

The procedure [RBH] (rescaled-BH) is defined as the step-up procedure using the critical values $\tau_k = \lambda_\alpha k/m$, $1 \leq k \leq m$, where $\lambda_\alpha = \max\{\lambda \in [0,1] : \Psi(\lambda_\alpha) \leq \alpha\}$ for

$$\Psi(\lambda) = \min\left(\lambda, \max_{1\leq k\leq m}\left(\frac{1}{k}\sum_{i=1}^{m}\frac{F_i(\lambda k/m)}{1-F_i(\lambda)}\right)\right).$$

The following result is straightforward from Theorem 1 (SU part).

Corollary 2. *In the setting of Theorem 1 with the additional assumption (2), we have* $\forall P \in \mathcal{P}$, $\mathrm{FDR}(\mathrm{RBH}, P) \leq \alpha$.

Moreover, if α is such that the equality $\Psi(\lambda_\alpha) = \alpha$ holds true, then $\lambda_\alpha \geq \Psi(\lambda_\alpha) = \alpha$ and [RBH] always dominates [BH] in terms of critical values and therefore rejects at least as many hypotheses.

A.2. A heterogeneous BR procedure

The procedure [HBR-λ] (discrete BR) is defined as the step-up procedure $\mathbf{SU}(\tau)$ using the critical values defined in the following way: for $k \in \{1, \ldots, m\}$,

$$\tau_k = \max\Big\{t \in \mathcal{A} \,:\, (F(t))_{(1)} \leq \lambda, \\ (F(t))_{(1)} + \cdots + (F(t))_{(m-k+1)} \leq \alpha k(1-\lambda)\Big\},$$

where each $(F(t))_{(j)}$ denotes the j-th largest elements of the set $\{F_i(t), 1 \leq i \leq m\}$. The following result is straightforward from Theorem 1 (SU part).

Corollary 3. *In the setting of Theorem 1, with the additional assumption (2), we have $\forall P \in \mathcal{P}$, $\mathrm{FDR}(\mathrm{HBR}, P) \leq \alpha$. Moreover, the set of nulls rejected by [HBR-λ] is larger than the one of [BR-λ] (almost surely), with equality (almost surely) under (4) and $F_i = F_j$ for all $i \neq j$.*

Appendix B: Supplement

B.1. Counterexample

We present here a modification of the counterexample due to Krieger given in Heller and Gur (2011). Consider $m = 3$ p-value null distributions given by

$$P_1 = 0.05 \cdot \delta_{\{0.05\}} + 0.16 \cdot \delta_{\{0.21\}} + 0.79 \cdot \delta_{\{1\}};$$
$$P_2 = 0.2 \cdot \delta_{\{0.2\}} + 0.09 \cdot \delta_{\{0.29\}} + 0.71 \cdot \delta_{\{1\}};$$
$$P_3 = \delta_{\{1\}},$$

where $\delta_{\{x\}}$ denotes the Dirac distribution in x. It is easy to verify that (1) yields

$$\overline{F}(t) = \begin{cases} 0 & t < 0.05; \\ 0.05/3 & t \in [0.05, 0.2); \\ 0.25/3 & t \in [0.2, 0.21); \\ 0.41/3 & t \in [0.21, 0.29); \\ 0.50/3 & t \in [0.29, 1); \\ 1 & t \geq 1. \end{cases}$$

Then the critical values of [Heyse] at level $\alpha = 0.25$ are given by $\tau_1 = 0.2$, $\tau_2 = \tau_3 = 0.29$, see (8). Now consider an alternative distribution for P_3 given by

$$Q_3 = \epsilon\delta_{\{0\}} + (1-\epsilon)\delta_{\{0.3\}},$$

where ϵ will be suitably chosen further on. Assume that the p-values p_1, p_2, p_3 are independent, with $p_i \sim P_i$ for $i \in \{1, 2\}$ (hypotheses H_1 and H_2 true) and $p_3 \sim Q_3$ (hypothesis H_3 false).

On the one hand, let us focus on the event $E = \{p_3 = 0.3\}$. In this case, H_3 is never rejected and the FDP of [Heyse] is 0 if and only if H_1 and H_2 are both not rejected and is equal to 1 otherwise. We partition E into the following different events:

- $E \cap \{p_1 = 0.05\}$: in this case, $p_{(1)} = 0.05 \leq \tau_1$ and at least H_1 will be (falsely) rejected and FDP = 1;
- $E \cap \{p_1 = 0.21, p_2 \neq 1\}$: in this case, $p_{(1)} = 0.2 \leq \tau_1$ and at least H_2 will be (falsely) rejected and FDP = 1;
- $E \cap \{p_1 = 0.21, p_2 = 1\}$: in this case, $p_{(1)} = 0.21 > \tau_1$ and $p_{(2)} = 0.3 > \tau_2$ so H_1 and H_2 are not rejected and FDP = 0;
- $E \cap \{p_1 = 1, p_2 = 0.2\}$: in this case, $p_{(1)} = 0.2 \leq \tau_1$ and H_2 will be (falsely) rejected and $FDP = 1$;
- $E \cap \{p_1 = 1, p_2 \neq 0.2\}$: in this case, $p_{(1)} = 0.29 > \tau_1$ and $p_{(2)} = 0.3 > \tau_2$ so H_1 and H_2 are not rejected and FDP = 0.

Altogether, we obtain

$$\mathbf{E}(\text{FDP} \times \mathbf{1}\{E\}) = (1 - \epsilon)(0.05 + 0.16 \times 0.29 + 0.79 \times 0.2) = (1 - \epsilon)0.2544.$$

On the other hand, let us focus on the event $E^c = \{p_3 = 0\}$. In this case, H_3 is always rejected and the FDP of [Heyse] can be 1/2 if one null is rejected among H_1 and H_2, and 2/3 if both H_1 and H_2 are rejected (it is 0 if both H_1 and H_2 are non rejected). We partition E into the following different events:

- $E^c \cap \{p_1 \neq 1, p_2 \neq 1\}$: in this case, $p_{(3)} \leq 0.29 = \tau_3$ and both H_1 and H_2 are rejected and FDP = 2/3;
- $E^c \cap \{p_1 \neq 1, p_2 = 1\}$: in this case, $p_{(2)} \leq 0.29 = \tau_2$ and $p_{(3)} = 1 > \tau_3$ so H_1 is rejected and not H_2. So FDP = 1/2;
- $E^c \cap \{p_1 = 1, p_2 \neq 1\}$: in this case, $p_{(2)} \leq 0.29 = \tau_2$ and $p_{(3)} = 1 > \tau_3$ so H_2 is rejected and not H_1. So FDP = 1/2;
- $E^c \cap \{p_1 = 1, p_2 = 1\}$: in this case only H_3 is rejected and FDP = 0.

Altogether, we obtain

$$\begin{aligned}\mathbf{E}(\text{FDP} \times \mathbf{1}\{E^c\}) &= \epsilon((2/3) \times 0.21 \times 0.29 + (1/2) \times (0.21 \times 0.71 + 0.79 \times 0.29)) \\ &= \epsilon\, 0.2297.\end{aligned}$$

Finally, we get

$$\text{FDR} = \epsilon\, 0.2297 + (1 - \epsilon)0.2544 = 0.25193 > \alpha,$$

by choosing $\epsilon = 0.1$.

B.2. Proofs for lemmas comparing procedures

The lemmas presented here rely on the fact that, there is almost surely no p-value in $[0,1]\backslash\mathcal{A}$ (both in the continuous and discrete cases). All symbols "$=$" or "$\subset$" are intended to be valid almost surely in this section.

A result which will be extensively used in the proofs of this section is the following one : for p-values valued in the set $\mathcal{A}$, then the step-up procedure with critical values τ_k, $1 \leq k \leq m$, has the same rejection set as the step-up procedure with critical values $\xi_k = \max\{t \in \mathcal{A} \,:\, t \leq \tau_k\}$, $1 \leq k \leq m$. This fact comes from the simple following observation: for all k,

$$\begin{aligned}\{1 \leq i \leq m \,:\, p_i \leq \tau_k\} &= \{1 \leq i \leq m \,:\, p_i \in \mathcal{A},\, p_i \leq \tau_k\} \\ &= \{1 \leq i \leq m \,:\, p_i \in \mathcal{A},\, p_i \leq \xi_k\} \\ &= \{1 \leq i \leq m \,:\, p_i \leq \xi_k\}.\end{aligned}$$

The ξ_k's are called the "effective" critical values of $\mathbf{SD}(\tau)$ or $\mathbf{SU}(\tau)$ in the sequel.

B.2.1. Proof of Lemma 1

The effective critical values of the BH procedure are given by the quantities $\xi_k = \max\{t \in \mathcal{A} \,:\, t \leq \alpha k/m\}$, $1 \leq k \leq m$. If (2) holds, then $\overline{F}(t) \leq t$ and each ξ_k is clearly smaller than the k-th critical values of [Heyse]. This implies that the rejection set of [Heyse] is larger than the one of [BH]. Conversely, under (4) and if $F_i = F_j = \overline{F}$ for all $i \neq j$, we always have $\overline{F}(t) = F_i(t) = t$ for $t \in \mathcal{A}$. This implies that the ξ_k's are the critical values of [Heyse] and shows the reversed inclusion.

B.2.2. Proof of Lemmas 2 and 3

Let τ_k, $1 \leq k \leq m$, be the critical values of [HSU]. Let us consider $\xi_k = \max\left\{t \in \mathcal{A} \,:\, t \leq \frac{\alpha}{1+\alpha}\frac{k}{m}\right\}$ the effective critical values of the [BH] procedure at level $\alpha/(1+\alpha)$. Now, for all $t \in [0,1]$, we have by (2),

$$\begin{aligned}\overline{F}_{\mathrm{SU}}(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{F_i(t)}{1-F_i(\tau_m)} &\leq \frac{t}{m}\sum_{i=1}^{m}\frac{1}{1-F_i(\tau_m)} = t\cdot(1+\overline{F}_{\mathrm{SU}}(\tau_m)) \\ &\leq t\cdot(1+\alpha), \qquad (23)\end{aligned}$$

where the last inequality follows from the definition of τ_m. Thus we have $\overline{F}_{\mathrm{SU}}(\xi_m) \leq \alpha$, which in turn implies $\xi_m \leq \tau_m$. Additionally, the bound (23) yields

for $1 \leq k < m$

$$\begin{aligned}\tau_k &= \max\left\{t \in \mathcal{A} \,:\, t \leq \tau_m,\, \overline{F}_{\mathrm{SU}}(t) \leq \alpha k/m\right\}\\ &\geq \max\left\{t \in \mathcal{A} \,:\, t \leq \tau_m,\, t(1+\alpha) \leq \alpha k/m\right\}\\ &= \max\left\{t \in \mathcal{A} \,:\, t(1+\alpha) \leq \alpha k/m\right\}\\ &= \xi_k,\end{aligned}$$

where we used that $\xi_m \leq \tau_m$. This proves Lemma 2. The proof of Lemma 3 is analogue and is left to the reader.

B.2.3. Proof of Lemma 4

Let us first focus on the case (i) and denote by τ_k, $1 \leq k \leq m$, the critical values of [AHSU]. From (2), we have for $1 \leq k \leq m-1$,

$$\begin{aligned}\tau_k &\geq \max\left\{t \in \mathcal{A} \,:\, t \leq \tau_m,\, t \leq \alpha k(1-\tau_m)/(m-k+1)\right\}\\ &= \max\left\{t \in \mathcal{A} \,:\, t \leq \left((1-\tau_m)\frac{\alpha k}{m-k+1}\right) \wedge \tau_m\right\},\end{aligned}$$

which correspond to the effective critical values of [BR-λ] with $\lambda = \tau_m$. Now consider the case (ii) and denote again by τ_k, $1 \leq k \leq m$, the critical values of [AHSD]. From (2), we have for $1 \leq k \leq m$,

$$\begin{aligned}\tau_k &\geq \max\left\{t \in \mathcal{A} \,:\, (m-k+1)t/(1-t) \leq \alpha k\right\}\\ &= \max\left\{t \in \mathcal{A} \,:\, t \leq \alpha k/(m-k(1-\alpha)+1)\right\}\end{aligned}$$

which correspond to the effective critical values of [GBS]. This implies the result.

B.3. Proofs of technical lemmas for step-down and step-up procedures

B.3.1. Proof of Lemma 5

First note that for any step-up procedure

$$\hat{k} = \max\left\{k \in \{0,1,\ldots,m\} \,:\, \sum_{i=1}^{m} \mathbf{1}\{p_i \leq \tau_k\} \geq k\right\},$$

which is sometimes more handy, because this definition avoids to rely explicitly on the order statistics of the p-values.

Now, it is not difficult to check that $\hat{k}^{\sharp,-i} \geq \hat{k} - 1$ always holds: this comes from the inequality

$$\hat{k} - 1 = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\hat{k}}\} - 1 \leq \sum_{j \neq i} \mathbf{1}\{p_j \leq \tau_{\hat{k}}\} = \sum_{j \neq i} \mathbf{1}\{p_j \leq \tau_{\hat{k}-1}^{\sharp,-i}\},$$

because $\tau_{\ell-1}^{\sharp,-i} = \tau_\ell$ for $\ell \in \{2,\dots,m\}$ (note that we can assume without loss of generality $\hat{k} \geq 1$ here). This means that (i) implies (ii). Now, when $p_i \leq \tau_{\hat{k}^{\sharp,-i}+1}$, we have

$$\hat{k}^{\sharp,-i} = \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp,-i}_{\hat{k}^{\sharp,-i}}\} = \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau_{\hat{k}^{\sharp,-i}+1}\} = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\hat{k}^{\sharp,-i}+1}\} - 1$$

which implies $\hat{k}^{\sharp,-i} + 1 = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\hat{k}^{\sharp,-i}+1}\}$ and thus $\hat{k}^{\sharp,-i} + 1 \leq \hat{k}$ (by using the definition of $\hat{k}$). Since, again, $\hat{k}^{\sharp,-i} \geq \hat{k} - 1$ always holds, we have $\hat{k}^{\sharp,-i} + 1 = \hat{k}$. Hence, (ii) implies (iii). Now, if $\hat{k}^{\sharp,-i} + 1 = \hat{k}$, we have

$$\begin{aligned}\mathbf{1}\{p_i \leq \tau_{\hat{k}}\} &= \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\hat{k}}\} - \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau_{\hat{k}}\} = \hat{k} - \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau_{\hat{k}^{\sharp,-i}+1}\} \\ &= \hat{k} - \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp,-i}_{\hat{k}^{\sharp,-i}}\} = \hat{k} - \hat{k}^{\sharp,-i} = 1,\end{aligned}$$

by definition of $\tau^{\sharp,-i}$, which gives that (iii) implies (i). Now, to prove the last statement, we first note that $\hat{k}^{\sharp} \geq \hat{k}^{\sharp,-i}$ always holds. Furthermore, if $p_i > \tau_m$ let us prove $\hat{k}^{\sharp} \leq \hat{k}^{\sharp,-i}$. First, $\hat{k}^{\sharp} = m$ is impossible because p_i is above τ_m and thus p_i cannot be rejected by $\mathbf{SU}^{\sharp}(\tau)$. Hence, $\hat{k}^{\sharp} \leq m-1$ and thus $\tau^{\sharp,-i}_{\hat{k}^{\sharp}}$ is well defined. Now, since $p_i > \tau_m$, we obtain

$$\sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp,-i}_{\hat{k}^{\sharp}}\} = \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp}_{\hat{k}^{\sharp}}\} = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau^{\sharp}_{\hat{k}^{\sharp}}\} = \hat{k}^{\sharp},$$

which implies $\hat{k}^{\sharp} \leq \hat{k}^{\sharp,-i}$ by definition of $\mathbf{SU}^{\sharp,-i}(\tau)$.

B.3.2. Proof of Lemma 6

First note that for any step-down procedure

$$\tilde{k} = \max\left\{k \in \{0,1,...,m\} \,:\, \forall k' \leq k,\ \sum_{i=1}^{m} \mathbf{1}\{p_i \leq \tau_{k'}\} \geq k'\right\}.$$

Now, we check that $\tilde{k}^{\sharp,-i} + 1 \geq \tilde{k}$ always holds. Since $\sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp,-i}_{\tilde{k}^{\sharp,-i}+1}\} < \tilde{k}^{\sharp,-i} + 1$, we have

$$\sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\tilde{k}^{\sharp,-i}+2}\} \leq 1 + \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp,-i}_{\tilde{k}^{\sharp,-i}+1}\} < \tilde{k}^{\sharp,-i} + 2,$$

which gives $\tilde{k} < \tilde{k}^{\sharp,-i} + 2$ by definition of $\tilde{k}$ and thus $\tilde{k} \leq \tilde{k}^{\sharp,-i} + 1$. Next, if $p_i \leq \tau_{\tilde{k}}$, we have

$$\sum_{j\neq i} \mathbf{1}\{p_j \leq \tau^{\sharp,-i}_{\tilde{k}}\} = \sum_{j\neq i} \mathbf{1}\{p_j \leq \tau_{\tilde{k}+1}\} = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\tilde{k}+1}\} - 1 < \tilde{k} + 1 - 1,$$

so that $\tilde{k} > \tilde{k}^{\sharp,-i}$ and thus $\tilde{k} \geq \tilde{k}^{\sharp,-i} + 1$. This proves that (i) implies (iv). Next, if $p_i > \tau_{\tilde{k}^{-i}+1}$, then

$$\sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\tilde{k}^{-i}+1}\} = \sum_{j \neq i} \mathbf{1}\{p_j \leq \tau_{\tilde{k}^{-i}+1}\} = \sum_{j \neq i} \mathbf{1}\{p_j \leq \tau^{-i}_{\tilde{k}^{-i}+1}\} < \tilde{k}^{-i} + 1,$$

which entails $\tilde{k} < \tilde{k}^{-i} + 1$ and thus $\tilde{k} \leq \tilde{k}^{-i}$. This proves $\tilde{k} \neq \tilde{k}^{\sharp,-i} + 1$. Hence, (iv) implies (iii). The fact that (iii) implies (ii) is obvious because $\tilde{k} \geq \tilde{k}^{-i}$ always holds. Finally, we merely check that $\tilde{k}$ is such that

$$\tilde{k} = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\tilde{k}}\} = \sum_{j=1}^{m} \mathbf{1}\{p_j \leq \tau_{\tilde{k}+1}\},$$

which means that the set of p-values rejected at threshold $\tau_{\tilde{k}}$ is the same as the set of p-values rejected at threshold $\tau_{\tilde{k}+1}$. This gives that (ii) implies (i). For the last assertion, it has been proved in the above reasoning while showing that (iv) implies (iii).

Appendix C: Empirical analyses for randomised *p*-values

In this section we follow the suggestion of one of the reviewers to investigate how using randomised p-values (see, e.g., Habiger, 2015) compares to our procedures. We do this by reanalysing the Pharmacovigilance and Arabidopsis methylation data from Section 5. To be more specific, we apply the BH and the Storey procedure (with $\lambda = 1/2$) to randomised p-values and denote these procedures by [r-BH] and [r-Storey]. For each random set of randomised p-values this results in a random set of rejected hypotheses. We repeat this simulation 1000 times and for each simulation run determine the number of rejected hypotheses. The resulting distribution of the number of rejected hypotheses is summarised numerically in Table 2 and displayed visually in Figure 4.

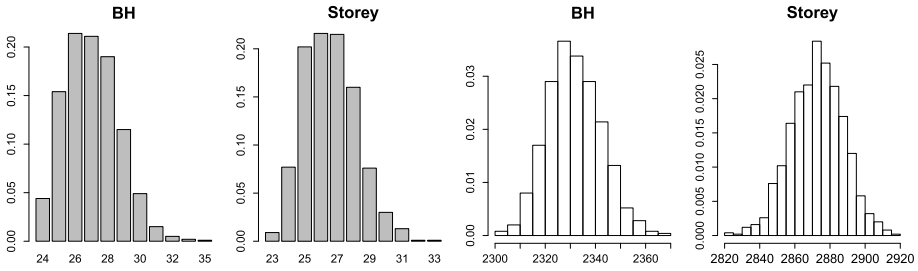
TABLE 2
Numerical summaries of rejections by randomised procedures.

Data set	Procedure	Min	1st Qu.	Median	Mean	3rd Qu.	Max
Pharmacovigilance	[r-BH]	24	26	27	27.02	28	35
	[r-Storey]	23	25	26	26.58	28	33
Arabidopsis methylation	[r-BH]	2302	2324	2331	2332	2339	2379
	[r-Storey]	2820	2863	2873	2873	2884	2916

The discrete BH procedure [HSU] compares favorably with [r-BH]: for the pharmacovigilance data the number of rejections by [HSU] (=27, see Table 1) is just the median of the distribution of [r-BH] and for the arabidopsis methylation data the number of rejections by [HSU] (=2358, see Table 1) is in the very right tail of the distribution of [r-BH].

The pharmacovigilance data seems to contain very few signals, so there is no benefit in using (either randomised or non-randomised) adaptive procedures

as compared to discrete procedures (in fact, [r-BH] is more powerful than [r-Storey]). This is also consistent with the findings in Sections 5 and 6. In contrast, the arabidopsis methylation data seems to contain a large portion of signals, so that adaptive procedures become effective. We see that [r-Storey] considerably outperforms the adaptive discrete procedures from Table 1 which are not based on a plug-in method. We think the reason for this phenomenon is that this seems to be a situation which is tailored to the strengths of the plug-in method (again this is consistent with the findings in Sections 5 and 6).



(a) Pharmacovigilance data (b) Arabidopsis methylation data

FIG 4. *Distribution of the number of rejected hypotheses when using randomised p-values.*

Summarizing the findings from this section, it appears that the amount of power that is lost by avoiding randomisation depends primarily on the proportion of alternatives. The Pharmacovigilance data set may serve as an example for a small proportion of alternatives. In this setting, no power is lost – on average – by using discreteness and avoiding randomisation. When the proportion of alternatives is large however (the Arabidopsis methylation data may be considered a prototype example here) we think that the behaviour of the randomised/discrete procedure is determined primarily by how the quantity of signal is estimated. To support this fact, we propose in Appendix D a new procedure that combines our approach with the Storey estimator which rejects the same order of hypotheses as [r-Storey].

Appendix D: A plug-in version of [HSU]

In this section, we sketch a discrete plug-in procedure in the spirit of Storey et al. (2004) for adapting to the unknown quantity of a discrete signal. To keep the exposition short, we describe this approach only for step-up procedures, however our ideas carry over directly to step-down procedures. As the proof of Theorem 1 shows, we have from (22) the following bound for FDR

$$\mathrm{FDR}(\mathbf{SU}(\tau), P) \leq \max_{1\leq k\leq m} \max_{\substack{A\subset\{1,\dots,m\}\\ |A|=m_0}} \left(\frac{1}{k} \sum_{i\in A} \frac{F_i(\tau_k)}{1-F_i(\tau_m)} \right) \tag{24}$$

where $m_0 = |\mathcal{H}_0|$ is the (unknown) number of true null hypotheses. Choosing critical value sequence $\tau_1(m_0), \ldots, \tau_m(m_0)$ that satisfy

$$\max_{\substack{A\subset\{1,\ldots,m\}\\|A|=m_0}} \sum_{i\in A} \frac{F_i\left(\tau_k(m_0)\right)}{1-F_i(\tau_m(m_0))} \leq k\cdot\alpha \tag{25}$$

yields a new HSU-type procedure which is adapted to the number of null hypotheses. In applications, m_0 is an unknown quantity which has to be estimated appropriately, for more details on this issue, see, e.g., Blanchard and Roquain (2009), Storey et al. (2004), Liang and Nettleton (2012), Heesen and Janssen (2016) and references therein. Our plug-in method works as follows:

1. Given the data, determine an appropriate estimate $\widehat{m}_0$ for m_0.
2. Apply the step-up procedure with critical values $\tau_1(\widehat{m}_0), \ldots, \tau_m(\widehat{m}_0)$.

We emphasize that this approach is only a heuristic one and currently we do not have a proof for FDR control.

Depending on the amount of signals and discreteness of p-values this approach can lead to strongly enhanced rejection numbers. As an example, we revisit the analysis of the Arabidopsis methylation data (see Section 5). Figure 5 depicts the number of rejections R as a function of $\widehat{\pi}_0 = \widehat{m}_0/m$ for this data set. The estimator used by the [Storey] procedure in Table 1 yields $\widehat{\pi}_0 = 0.6$ and thus the corresponding discrete plug-in procedure rejects $R = 2659$ hypotheses. The randomised p-values that were used for evaluating [r-Storey] in Appendix C result in an average $\widehat{\pi}_0 = 0.468$. Using this estimate, the discrete plug-in procedure rejects $R = 2836$ hypotheses, which lies within the range of the rejection numbers for the completely randomised procedure [r-Storey].

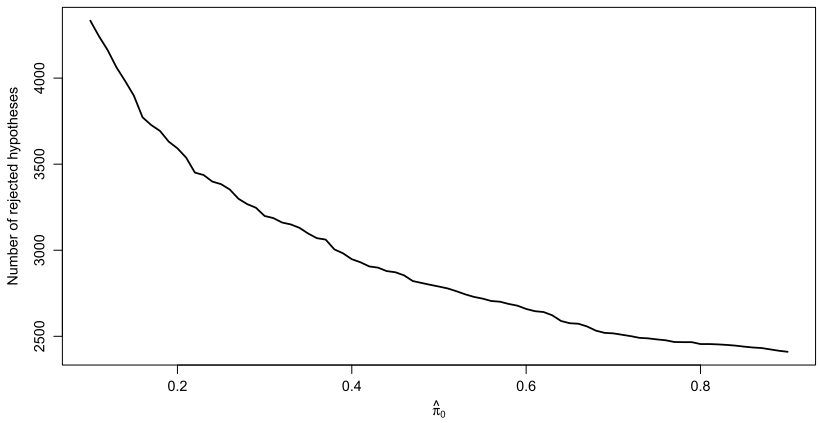


FIG 5. *Number of rejections of the discrete plug-in procedure for the arabidopsis methylation data.*

Appendix E: Tables for the simulations

TABLE 3
Average power in the simulation study (see Section 6).

m	m_3	m_1	q	[BH]	[Storey]	[Heyse]	[HSU]	[AHSU]
800	80	144	0.15	0.0000	0.0000	0.0004	0.0003	0.0003
		144	0.25	0.0004	0.0004	0.0197	0.0183	0.0183
		144	0.4	0.0803	0.0559	0.4425	0.4268	0.4268
		360	0.15	0.0000	0.0000	0.0007	0.0006	0.0006
		360	0.25	0.0004	0.0003	0.0244	0.0221	0.0221
		360	0.4	0.0803	0.0514	0.4529	0.4518	0.4518
		576	0.15	0.0000	0.0000	0.0009	0.0008	0.0008
		576	0.25	0.0004	0.0003	0.0343	0.0278	0.0278
		576	0.4	0.0803	0.0484	0.5367	0.4832	0.4832
	240	112	0.15	0.0000	0.0000	0.0003	0.0003	0.0003
		112	0.25	0.0005	0.0004	0.0276	0.0257	0.0257
		112	0.4	0.2148	0.1963	0.5365	0.5152	0.5152
		280	0.15	0.0000	0.0000	0.0003	0.0003	0.0003
		280	0.25	0.0005	0.0004	0.0315	0.0282	0.0282
		280	0.4	0.2147	0.1883	0.5758	0.5596	0.5596
		448	0.15	0.0000	0.0000	0.0005	0.0004	0.0004
		448	0.25	0.0005	0.0003	0.0372	0.0323	0.0323
		448	0.4	0.2145	0.1793	0.5920	0.5844	0.5844
	640	32	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		32	0.25	0.0010	0.0016	0.0378	0.0352	0.0352
		32	0.4	0.4243	0.6519	0.6174	0.5983	0.6838
		80	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		80	0.25	0.0010	0.0014	0.0388	0.0359	0.0359
		80	0.4	0.4242	0.6370	0.6282	0.6146	0.6848
		128	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		128	0.25	0.0010	0.0013	0.0400	0.0368	0.0368
		128	0.4	0.4240	0.6276	0.6353	0.6271	0.6859
2000	200	360	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		360	0.25	0.0001	0.0001	0.0156	0.0145	0.0145
		360	0.4	0.0730	0.0499	0.4486	0.4334	0.4334
		900	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		900	0.25	0.0001	0.0001	0.0192	0.0170	0.0170
		900	0.4	0.0730	0.0439	0.4517	0.4517	0.4517
		1440	0.15	0.0000	0.0000	0.0003	0.0003	0.0003
		1440	0.25	0.0001	0.0001	0.0286	0.0218	0.0218
		1440	0.4	0.0730	0.0402	0.5402	0.4748	0.4748
	600	280	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		280	0.25	0.0001	0.0001	0.0239	0.0217	0.0217
		280	0.4	0.2058	0.1953	0.5350	0.5166	0.5166
		700	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		700	0.25	0.0001	0.0001	0.0290	0.0246	0.0246
		700	0.4	0.2058	0.1917	0.5750	0.5630	0.5630
		1120	0.15	0.0000	0.0000	0.0002	0.0002	0.0002
		1120	0.25	0.0001	0.0001	0.0350	0.0296	0.0296
		1120	0.4	0.2057	0.1832	0.5908	0.5853	0.5853
	1600	80	0.15	0.0000	0.0000	0.0001	0.0001	0.0001
		80	0.25	0.0003	0.0007	0.0379	0.0352	0.0352
		80	0.4	0.4223	0.6498	0.6196	0.5942	0.6863
		200	0.15	0.0000	0.0000	0.0001	0.0001	0.0001
		200	0.25	0.0003	0.0006	0.0387	0.0361	0.0361
		200	0.4	0.4222	0.6352	0.6281	0.6157	0.6871
		320	0.15	0.0000	0.0000	0.0001	0.0001	0.0001
		320	0.25	0.0003	0.0005	0.0396	0.0369	0.0369
		320	0.4	0.4220	0.6282	0.6327	0.6279	0.6880

TABLE 4
Average FDR in the simulation study (see Section 6).

m	m_3	m_1	q	[BH]	[Storey]	[Heyse]	[HSU]	[AHSU]
800	80	144	0.15	0.000000	0.000000	0.000030	0.000021	0.000021
		144	0.25	0.000000	0.000000	0.000076	0.000066	0.000066
		144	0.4	0.000005	0.000002	0.001228	0.001154	0.001154
		360	0.15	0.000000	0.000000	0.000035	0.000030	0.000030
		360	0.25	0.000000	0.000000	0.000081	0.000067	0.000067
		360	0.4	0.000004	0.000001	0.000823	0.000797	0.000797
		576	0.15	0.000000	0.000000	0.000020	0.000017	0.000017
		576	0.25	0.000000	0.000000	0.000061	0.000042	0.000042
		576	0.4	0.000002	0.000001	0.001148	0.000915	0.000915
	240	112	0.15	0.000000	0.000000	0.000021	0.000021	0.000021
		112	0.25	0.000000	0.000000	0.000159	0.000139	0.000139
		112	0.4	0.000101	0.000062	0.004636	0.004540	0.004540
		280	0.15	0.000000	0.000000	0.000014	0.000013	0.000013
		280	0.25	0.000000	0.000000	0.000130	0.000106	0.000106
		280	0.4	0.000063	0.000032	0.003226	0.002962	0.002962
		448	0.15	0.000000	0.000000	0.000010	0.000007	0.000007
		448	0.25	0.000000	0.000000	0.000080	0.000060	0.000060
		448	0.4	0.000025	0.000012	0.002606	0.001583	0.001583
	640	32	0.15	0.000000	0.000000	0.000012	0.000012	0.000012
		32	0.25	0.000000	0.000001	0.000308	0.000252	0.000252
		32	0.4	0.001253	0.014708	0.014557	0.014527	0.015222
		80	0.15	0.000000	0.000000	0.000011	0.000011	0.000011
		80	0.25	0.000000	0.000000	0.000218	0.000176	0.000176
		80	0.4	0.000793	0.009106	0.009118	0.009111	0.009542
		128	0.15	0.000000	0.000000	0.000005	0.000005	0.000005
		128	0.25	0.000000	0.000000	0.000092	0.000071	0.000071
		128	0.4	0.000323	0.003566	0.003654	0.003653	0.003846
2000	200	360	0.15	0.000000	0.000000	0.000011	0.000011	0.000011
		360	0.25	0.000000	0.000000	0.000043	0.000038	0.000038
		360	0.4	0.000003	0.000001	0.001251	0.001197	0.001197
		900	0.15	0.000000	0.000000	0.000010	0.000008	0.000008
		900	0.25	0.000000	0.000000	0.000045	0.000035	0.000035
		900	0.4	0.000002	0.000001	0.000790	0.000790	0.000790
		1440	0.15	0.000000	0.000000	0.000005	0.000004	0.000004
		1440	0.25	0.000000	0.000000	0.000041	0.000025	0.000025
		1440	0.4	0.000001	0.000000	0.001160	0.000962	0.000962
	600	280	0.15	0.000000	0.000000	0.000011	0.000010	0.000010
		280	0.25	0.000000	0.000000	0.000115	0.000093	0.000093
		280	0.4	0.000068	0.000056	0.004615	0.004571	0.004571
		700	0.15	0.000000	0.000000	0.000007	0.000007	0.000007
		700	0.25	0.000000	0.000000	0.000105	0.000081	0.000081
		700	0.4	0.000041	0.000033	0.003076	0.002969	0.002969
		1120	0.15	0.000000	0.000000	0.000003	0.000003	0.000003
		1120	0.25	0.000000	0.000000	0.000057	0.000045	0.000045
		1120	0.4	0.000016	0.000012	0.002569	0.001592	0.001592
	1600	80	0.15	0.000000	0.000000	0.000004	0.000004	0.000004
		80	0.25	0.000000	0.000001	0.000256	0.000229	0.000229
		80	0.4	0.001226	0.014589	0.014515	0.014499	0.015228
		200	0.15	0.000000	0.000000	0.000002	0.000002	0.000002
		200	0.25	0.000000	0.000000	0.000173	0.000152	0.000152
		200	0.4	0.000768	0.009109	0.009108	0.009105	0.009563
		320	0.15	0.000000	0.000000	0.000001	0.000001	0.000001
		320	0.25	0.000000	0.000000	0.000073	0.000061	0.000061
		320	0.4	0.000305	0.003641	0.003646	0.003646	0.003830

Acknowledgements

We would like to thank two anonymous referees and an associate editor for their helpful comments. This work has been supported by the CNRS (PEPS FaSciDo) and the French grants ANR-16-CE40-0019 (SansSouci project) and ANR-17-CE40-0001 (Basics project). We would also like to thank Xiongzhi Chen and Thorsten Dickhaus for helpful discussions.

References

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B 57*(1), 289–300. MR1325392

Benjamini, Y., A. M. Krieger, and D. Yekutieli (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika 93*(3), 491–507. MR2261438

Benjamini, Y. and W. Liu (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *J. Statist. Plann. Inference 82*(1-2), 163–170. MR1736441

Berger, R. L. (1996). More powerful tests from confidence interval p values. *The American Statistician 50*(4), 314–318.

Blanchard, G., T. Dickhaus, E. Roquain, and F. Villers (2014). On least favorable configurations for step-up-down tests. *Statist. Sinica 24*(1), 1–23. MR3184590

Blanchard, G. and E. Roquain (2009). Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res. 10*, 2837–2871. MR2579914

Chen, X. and R. Doerge (2015a). *fdrDiscreteNull: False Discovery Rate Procedure Under Discrete Null Distributions*. R package version 1.0.

Chen, X. and R. Doerge (2015b). A weighted fdr procedure under discrete and heterogeneous null distributions. *arXiv:1502.00973*.

Dickhaus, T. (2014). *Simultaneous statistical inference*. Springer, Heidelberg. With applications in the life sciences. MR3184277

Dickhaus, T., K. Straßburger, D. Schunk, C. Morcillo-Suarez, T. Illig, and A. Navarro (2012). How to analyze many contingency tables simultaneously in genetic association studies. *Statistical applications in genetics and molecular biology 11*(4). MR2958611

Döhler, S. (2016). A discrete modification of the Benjamini—Yekutieli procedure. *Econometrics and Statistics*. MR3740116

Durand, G. (2017). Adaptive p-value weighting with power optimality. *arXiv:1710.01094*.

Ferreira, J. A. (2007). The Benjamini-Hochberg method in the case of discrete test statistics. *Int. J. Biostat. 3*, Art. 11, 18. MR2383611

Ferreira, J. A. and A. H. Zwinderman (2006). On the Benjamini-Hochberg method. *Ann. Statist. 34*(4), 1827–1849. MR2283719

Finner, H., T. Dickhaus, and M. Roters (2009). On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist. 37*(2), 596–618. MR2502644

Gavrilov, Y., Y. Benjamini, and S. K. Sarkar (2009). An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist. 37*(2), 619–629. MR2502645

Gilbert, P. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society. Series C 54*(1), 143–158. MR2134603

Habiger, J. D. (2015). Multiple test functions and adjusted p-values for test statistics with discrete distributions. *J. Statist. Plann. Inference 167*, 1–13. MR3383232

Heesen, P. and A. Janssen (2016). Dynamic adaptive multiple tests with finite sample fdr control. *Journal of Statistical Planning and Inference 168*, 38 – 51. MR3412220

Heller, R. and H. Gur (2011). False discovery rate controlling procedures for discrete tests. *arXiv:1112.4627*.

Heller, R., H. Gur, and S. Yaacoby (2012). *discreteMTP: Multiple testing procedures for discrete test statistics*. R package version 0.1-2.

Heyse, J. F. (2011). A false discovery rate procedure for categorical data. In *Recent Advances in Bio- statistics: False Discovery Rates, Survival Analysis, and Related Topics*, pp. 43–58.

Ignatiadis, N., B. Klaus, J. B. Zaugg, and W. Huber (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods 13*(7), 577.

Karp, N. A., R. Heller, S. Yaacoby, J. K. White, and Y. Benjamini (2016). Improving the identification of phenotypic abnormalities and sexual dimorphism in mice when studying rare event categorical characteristics. *Genetics*.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association 56*(294), 223–234. MR0124107

Liang, K. and D. Nettleton (2012). Adaptive and dynamic adaptive procedures for false discovery rate control and estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(1), 163–182. MR2885844

Lister, R., R. C. O'Malley, J. Tonti-Filippini, B. D. Gregory, C. C. Berry, A. H. Millar, and J. R. Ecker (2008, May). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell 133*(3), 523–536.

Mantel, N. (1980). A biometrics invited paper. assessing laboratory evidence for neoplastic activity. *Biometrics 36*(3), 381–399.

Pounds, S. and C. Cheng (2006). Robust estimation of the false discovery rate. *Bioinformatics 22*(16), 1979–1987.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ripamonti, E., C. Lloyd, and P. Quatto (2017). Contemporary frequentist views of the 2x2 binomial trial. *Statistical Science*. MR3730524

Roquain, E. and M. van de Wiel (2009). Optimal weighting for false discovery rate control. *Electron. J. Stat. 3*, 678–711. MR2521216

Roquain, E. and F. Villers (2011). Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist. 39*(1), 584–612. MR2797857

Storey, J. D., J. E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(1), 187–205. MR2035766

Tarone, R. E. (1990). A modified bonferroni method for discrete data. *Biometrics 46*(2), 515–522.

van den Broek, E., M. J. J. Dijkstra, O. Krijgsman, D. Sie, J. C. Haan, J. J. H. Traets, M. A. van de Wiel, I. D. Nagtegaal, C. J. A. Punt, B. Carvalho, B. Ylstra, S. Abeln, G. A. Meijer, and R. J. A. Fijneman (2015, 09). High prevalence and clinical relevance of genes affected by chromosomal breaks in colorectal cancer. *PLOS ONE 10*(9), 1–14.

Westfall, P. and R. Wolfinger (1997). Multiple tests with discrete distributions. *The American Statistician 51*(1), 3–8.