Stat Papers
https://doi.org/10.1007/s00362-018-1008-y

REGULAR ARTICLE

A comprehensive error rate for multiple testing

Djalel-Eddine Meskaldji[1,3,4] · Dimitri Van De Ville[2,3] ·
Jean-Philippe Thiran[4] · Stephan Morgenthaler[1]

Received: 15 June 2016 / Revised: 30 April 2018

Abstract Over the last two decades, a large variety of type I error rates and control procedures have been proposed in the field of multiple hypotheses testing. This paper proposes a framework that includes many existing proposals by investigating procedures in which the ordered *p*-values are compared to an arbitrary positive and non-decreasing threshold sequence. For this case, we derive the error rate being controlled under different assumptions on the *p*-values. Our focus will be on step-up procedures. The new formulation gives insight into the relations between existing error rates and opens new perspectives for the whole field of multiple testing.

Keywords Family-wise error rate · False discovery rate · Multiple comparisons · Ordered p-values · Scaled false discovery rate · Type I error control

1 Introduction

When testing m hypotheses, it is customary to compute an individual p-value p_i for each null hypothesis H_i, $i \in \{1, \ldots, m\}$. A true effect (false hypothesis) tends to have a p-value close to zero, while a null effect (true hypothesis) has a p-value uniformly

✉ Djalel-Eddine Meskaldji
djalel.meskaldji@epfl.ch

[1] Institute of Mathematics, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[2] Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

[3] Institute of Bioengineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[4] Institute of Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Published online: 18 May 2018

distributed between 0 and 1 or stochastically lower-bounded by a uniform variable $\mathcal{U}(0, 1)$, i.e.,

$$\mathbb{P}\left(p_i \leq u\right) \leq u \quad \text{for all } i \in \mathcal{H}_0, \tag{1}$$

where $\mathcal{H}_0 = \{i \in \{1, \ldots, m\} : H_i \text{ is true}\}$ is the set of the indices of the true null hypotheses. Let $\mathcal{H}_1 = \{1, \ldots, m\} \backslash \mathcal{H}_0$ be the set of the indices of the false null (alternative) hypotheses.

A multiple hypotheses testing procedure is a decision function that computes from the data whether or not each hypothesis is rejected, that is, a vector $\{R_1, \ldots, R_m\}$ where $R_i = 1$ if the i^{th} null hypothesis is rejected and 0 otherwise. The resulting rejection set $\mathcal{R}$ is the subset of $\{1, \ldots, m\}$ that contains the rejected hypotheses, that is, $\mathcal{R} = \{i \in \{1, \ldots, m\} : R_i = 1\}$. The number of erroneously rejected hypotheses, i.e., the number of false positives, is equal to the number of the rejected true hypotheses, i.e., $V = |\mathcal{R} \cap \mathcal{H}_0|$. Let $m = m_0 + m_1$ be the decomposition into the number of true null hypotheses and false hypotheses, respectively. Table 1 introduces the contingencies that occur.

A multiple testing procedure is designed to control a type I error rate under certain assumptions on the dependence of the p-values. In this paper, we consider independence, positive dependence and general dependence. In the general dependence case, only property (1) is required.

The best known multiple comparison procedure is the Bonferroni procedure which performs each of the m tests at significance level α/m or equivalently, declares an effect as significant, if its corresponding p-value is less than or equal to α/m. Based on the Bonferroni (1936) inequality, this procedure guarantees $\mathbb{E}(V) \leq \alpha$ under general dependence. This control implies the control of the family-wise error rate FWER $= \mathbb{P}(V > 0)$. Traditionally, this strong control was favored. Scientists who used the Bonferroni procedure noted, however, that its ability to find true effects vanishes when the number m of null hypotheses being tested becomes large.

The plot of the ordered p-values against the rank provides a simple visual check to see whether any effect is present at all (Bernhard et al. 2004). The resulting graph of $p_{(r)}$ vs r for $r = 1, \ldots, m$ carves out an non-decreasing sequence of points in the rectangle $[1; m] \times [0; 1]$ and should, if no real effect exists, stay close to the diagonal from the lower left corner to the upper right corner of the square (Benjamini 2010; Schweder and Spjøtvoll 1982).

Multiple testing procedures can be based on a second curve or threshold sequence t_r $(r = 1, \ldots, m)$, which stays close to the lower edge of the square. If all p-values

Table 1 General outcome of separating m tests into non significant and significant ones

Hypotheses	Not rejected	Rejected	Total
True	U	V	m_0
Non-true/false	T	S	$m - m_0 = m_1$
Total	$m - R$	R	m

S, V denote the number of true, false rejections (true, false positives), respectively, *R* is the total number of rejections, *U, T* denote the number of true, false non-rejections (true, false negatives), respectively

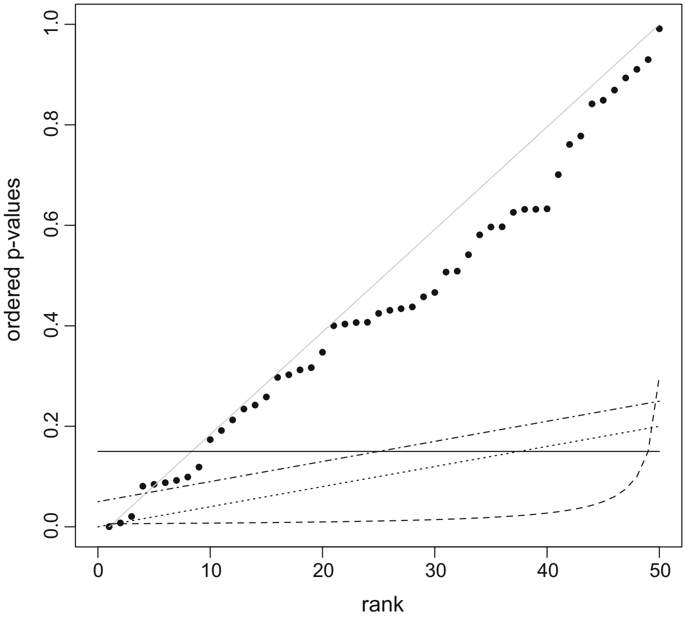

Fig. 1 The plot shows for $m = 50$ tests the ordered p-values plotted against the rank. Besides the diagonal, three other curves are shown, which have appeared in the literature as thresholds. They are the constant threshold $t_r = 0.15$ (solid line), the proportional threshold $t_r = 0.004r$ (dotted line) and the curved threshold $t_r = 0.3/(m + 1 - r)$ (dashed curve). With the constant, 9 tests reject, with the proportional, 1 test rejects and with the curved, 2 tests reject. In addition a linear threshold $t_r = 0.05 + 0.004r$ is shown

are above this curve, none of the tests rejects. If any of the p-values are below the threshold curve, there are several possible procedures. Examples of such procedures include Holm (1979), Simes (1986) and Hochberg (1988).

The step-up procedures look for the last up-crossing of the threshold by the ordered p-values. The following is a description of this technique:

1. choose a non-decreasing threshold sequence $t_1 \leq t_2 \leq \cdots \leq t_m$;
2. let r be the largest rank such that $p_{(r)} \leq t_r$ is true, that is, the largest ordered p-value below the threshold;
3. reject the null hypotheses corresponding to the ranks $1, \ldots, r$;
4. if all the p-values are above the threshold, no null hypothesis is rejected.

The rejection set of a step-up procedure is given by

$$\mathcal{R} = \left\{r \in \{1, \ldots, m\} : p_{(r)} \leq p_{(\tilde{r})}\right\},$$

where $p_{(\tilde{r})} = \max\left\{p_{(r)} \leq t_r\right\}$. Different tests are obtained by choosing different threshold sequences (Fig. 1). In all the examples given in Fig. 1, the values of the thresholds lie between α/m and α. This is reasonable since these two bounds represent the cases of "severe correction" and "no correction".

Benjamini and Hochberg (1995) proved that the step-up procedure with the linear threshold sequence $t_r = r\alpha/m, r = 1, \ldots, m$, guarantees the control of the False

Discovery Rate FDR $\equiv \mathbb{E}(V/R)$ (with FDR $= 0$ if $R = 0$) at level α if the tests are independent. Benjamini and Yekutieli (2001) extended the control to a certain positive dependence condition called positive regression dependency on each one from a subset (PRDS).

Since the introduction of the FDR by Benjamini and Hochberg (1995), the idea that the control of false positives V should be considered in conjunction with the number of rejections R has been widely accepted (Benjamini 2010).

One of the main objectives of this paper is to study this conjunction in more generality. Specifically, we explore what happens, if a step-up procedure is applied with an arbitrary non-decreasing threshold sequence t_r.

If we are willing to contemplate new error rates, the answer to "what happens with an arbitrary non-decreasing threshold sequence?" has an easy answer. We will show that, under independence and positive dependence, a step-up procedure with non-decreasing threshold sequence $t_r,\ r = 1, \ldots, m$, guarantees

$$\frac{1}{m}\mathbb{E}\left(\mathbf{1}\{R > 0\}\frac{V}{t_R}\right) \leq 1\,. \tag{2}$$

In order to relate this inequality to familiar results, we consider thresholds of the form

$$t_r = s(r)\alpha/m\,, \tag{3}$$

where s is a non-decreasing real function, which we will call the *shape function* when dealing with thresholds, and $\alpha > 0$ is an arbitrary tuning constant that controls the global severity of the error control. The bigger α, the larger the threshold and the larger the number of rejections.

Now, we start with the observation that the Bonferroni procedure can be considered as a step-up procedure with constant threshold $t_r \equiv \alpha/m$. If we put $s_{\mathrm{Bonf}}(r) \equiv 1,\ r = 1, \ldots, m$, as the Bonferroni shape function, then the step-up procedure with threshold sequence $t_r = s_{\mathrm{Bonf}}(r)\alpha/m,\ r = 1, \ldots, m$, controls $\mathbb{E}(V) = \mathbb{E}(V/s_{\mathrm{Bonf}}(R))$ at level α under general dependence. Likewise, if we set $s_{\mathrm{BH}}(r) \equiv r,\ r = 1, \ldots, m$, as the Benjamini and Hochberg shape function, the step-up procedure with threshold sequence $t_r = s_{\mathrm{BH}}(r)\alpha/m,\ r = 1, \ldots, m$, controls $\mathbb{E}(V/R) = \mathbb{E}(V/s_{\mathrm{BH}}(R))$ at level α under independence and positive dependence (PRDS). How general is this phenomenon with an arbitrary non-decreasing shape function s? Does it then hold that the expectation $\mathbb{E}\left[V/s(R)\right]$ is controlled by a step-up procedure with $t_r = s(r)\alpha/m,\ r = 1, \ldots, m$, and under which assumption? The inequality (2) is equivalent to

$$\mathbb{E}\left(\mathbf{1}\{R > 0\}\frac{V}{s(R)}\right) \leq \alpha\,. \tag{4}$$

That is, the function s acts as a *scale* function that moderates the influence of the number of rejections R in the error rate.

The matched error rate is given by the following definition.

Definition 1.1 We call the quantity

$$sFDP = \frac{V}{s(R)}\,,\ \text{if } s(R) > 0 \text{ and } = 0\,,\ \text{otherwise}\,,$$

Table 2 The control exerted by using a step-up procedure with the thresholds discussed in Sect. 1

Threshold t_r	Shape $s(r)$	Control	Assumption
α/m	1	$\mathbb{E}(1_{\{R>0\}}V) = \mathbb{E}(V) \le \alpha$	General dep.
$\alpha/(m+1-r)$	$\frac{m}{m+1-R}$	$\mathbb{E}(1_{\{R>0\}}\frac{m+1-R}{m}V) \le \alpha$	Indep. and pos. dep.
$r\alpha/m$	r	$\mathbb{E}(1_{\{R>0\}}\frac{V}{R}) \le \alpha$	Indep. and pos. dep.
$A+Br$	$\frac{m(A+Br)}{\alpha}$	$\mathbb{E}(1_{\{R>0\}}\frac{V}{m(A+BR)}) \le 1$	Indep. and pos. dep.

The first threshold (Bonferroni threshold) controls the error rate under any dependency structure. The second threshold sequence corresponds to one used in Holm step-down and Hochberg step-up procedures. In our framework, used in a step-up procedure, this threshold sequence controls (under independence and positive dependence) the expected number of erroneous rejections, but multiplied by a weight smaller or equal to 1 that decreases linearly as R increases. The linear $A+Br$, $(A, B) \neq (0, 0)$ and in particular, the proportional $r\alpha/m$ thresholds also control such a weighted expectation (under independence and positive dependence), with decreasing weights depending on R

the scaled false discovery proportion with scale function s. We call the expectation of the *sFDP* the scaled false discovery rate, sFDR $= \mathbb{E}\,(sFDP)$.

This appeared for the first time in print in Meskaldji et al. (2011).

For the thresholds we have discussed in Sect. 1, the inequality (4) yields the results in Table 2.

Our results are valid for $m \geq 2$ and cover the sFDR control via step-up procedures under independence, positive dependence and general dependence (with modified threshold sequences), as well as the control via weighted procedures (Benjamini and Hochberg 1997; Genovese et al. 2006). In a weighted procedure, an a priori non-negative weight w_i is assigned to each null hypothesis $H_i,\ i \in \{1, \dots, m\}$, and the raw p-values p_i are replaced by the weighted p-values $q_i = p_i/w_i$ if $w_i > 0$ and $q_i = 1$ if $w_i = 0$. A hypothesis with big weight is thus more easily rejected than a hypothesis with small weight. The weights have to be standardized so that $\sum_{i=1}^{m} w_i = m$, which includes $w_i \equiv 1, i = 1, \dots, m$, as a possibility.

The remainder of the paper is as follows. In Sect. 2 we formulate the results and give the proofs. In Sect. 3 we interpret the results, simulate multiple tests. Finally, we conclude in Sect. 4.

2 The control of the sFDR

The theory developed in Blanchard and Roquain (2008) can be adapted to our error rate. Three of their lemmas form the basis, together with the concept of *self consistency*.

Any step-up procedure is such that the rejection set satisfies $\mathcal{R} = \{i \in I = \{1, \dots, m\} : p_i \leq \Upsilon\}$, where Υ is an upper bound which usually depends on the p-values $p_1, \dots, p_m$. To prove control, we need to bound

$$\begin{aligned}\mathrm{sFDR}(\Upsilon) &= \mathbb{E}\left[\frac{V}{s(R)}\mathbf{1}\left\{R > 0\right\}\right] \\ &= \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\left\{p_i \leq \Upsilon\right\}}{s(R)}\mathbf{1}\left\{R > 0\right\}\right]. \end{aligned} \tag{5}$$

The family of threshold sequences $\Gamma : I \times \mathbb{R}^+ \mapsto \mathbb{R}^+$ of the form

$$\Gamma(i, r) = w_i s(r)\frac{\alpha}{m}\,,$$

where i refers to the ith null hypothesis and r denotes the rank of the p-value p_i, covers all cases of interest to us. It takes account of the weights $\{w_i, i \in I\}$ and the shape function $s(r)$. Such a threshold sequence leads to a step-up procedure that satisfies

$$\mathcal{R} \subset \{i \in I \mid p_i \leq \Gamma(i, R)\}\,.$$

This holds for all R and is called the self-consistency condition (Blanchard and Roquain 2008).

2.1 Control of the sFDR under independence

Our first result covers the independent case, in which

$$\{p_i : i \in \mathcal{H}_0\} \text{ are mutually independent and independent of } \{p_i : i \in \mathcal{H}_1\}. \tag{6}$$

Proposition 2.1 *Assume that $p_1, \ldots, p_m$ are independent p-values in the sense of (6) and that non-negative weights $w_1, \ldots, w_m$ are given. Then the step-up procedure with threshold sequence $t_r = s(r)\alpha/m$ applied to $q_1 = p_1/w_1, \ldots, q_m = p_m/w_m$ ($q_i = 1$ if $w_i = 0$), satisfies*

$$\mathrm{sFDR} \leq \alpha/m \sum_{i=1}^{m} w_i\,.$$

Proof Denote by $\mathbf{p}_{-i}$ the set of p-values $\{p_j \mid j \neq i\}$. From (5) and using the self-consistency, we have

$$\begin{aligned}\mathrm{sFDR}(R) &= \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\left\{p_i/w_i \leq s(R)\alpha/m\right\}}{s(R)}\right] \\ &\leq \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left\{p_i \leq \frac{\alpha}{m} w_i s(R(\mathbf{p}))\right\}}{s(R(\mathbf{p}))}\middle|\mathbf{p}_{-i}\right]\right] \\ &\leq \frac{\alpha}{m}\sum_{i\in\mathcal{H}_0} w_i\,,\end{aligned}$$

where we used the fact that given $\mathbf{p}_{-i}$, p_i is stochastically lower bounded by a uniform distribution in the independent case. By case (i) of the Lemma 3.2 of Blanchard and Roquain (2008), we obtain the result if we set $U = p_i$, $g(U) = s(R(\mathbf{p}_{-i}, U))$ and $c = \frac{\alpha}{m} w_i$. Note that $g(s(R(\mathbf{p}_{-i}, U)))$ is non increasing in U since s is non-decreasing and R is non-increasing in U. □

2.2 Control of the sFDR under positive dependence

Many positive dependencies have been considered in the literature such as MTP2 (Sarkar 1998) or martingale based dependency (Heesen and Janssen 2015).

The positive dependence that we consider in this paper is the one defined in Blanchard and Roquain (2008). It is based on the notion of non-decreasing subsets. A subset of the unit hypercube $D \subset [0,1]^m$ is a non-decreasing subset if for all $\mathbf{q}, \mathbf{q}' \in \mathbf{[0,1]^m}$ such that $\forall i \in \{1, \ldots, m\}, q_i \leq q_i'$, we have $\mathbf{q} \in D \implies \mathbf{q}' \in D$. The positive dependence condition is as follows.

$$\begin{gathered}\text{For any } i_0 \in \mathcal{H}_0 \text{ and any measurable non-decreasing set } D \subset [0,1]^m,\\ \text{the function} u \mapsto \mathbb{P}\big((p_i, 1 \leq i \leq m) \in D \mid p_{i_0} \leq u\big)\\ \text{is non-decreasing on the set } \{u \in [0,1] : \mathbb{P}(p_{i_0} \leq u) > 0\}.\end{gathered} \tag{7}$$

This condition is called weak positive regression dependency on each one from a subset (wPRDS), which is weaker than the PRDS defined in Benjamini and Yekutieli (2001), as pointed out by Roquain (2015) and Blanchard and Roquain (2008).

Proposition 2.2 *Assume that the p-values $p_1, \ldots, p_m$ are positively dependent random variables in the sense of wPRDS, and that non negative weights $w_1, \ldots, w_m$ are given. Then the step-up procedure with threshold sequence $t_r = s(r)\alpha/m$ applied to $q_1 = p_1/w_1, \ldots, q_m = p_m/w_m$ ($q_i = 1$ if $w_i = 0$), satisfies*

$$\text{sFDR} \leq \alpha/m \sum_{i=1}^{m} w_i \,.$$

Proof By case (ii) of the Lemma 3.2 of Blanchard and Roquain (2008), with $U = p_i$, $V = s(R)$ and $c = \frac{\alpha}{m} w_i$, the result follows. □

2.3 Control of the sFDR under general dependence

Under arbitrary dependence of the p-values, an additional correction to the thresholds has to be performed. Let ν be a probability distribution on $(0, \infty)$ and define for $r > 0$

$$\xi(r) = \int_0^r u d\nu(u)\,. \tag{8}$$

We call the ξ function *the reshaping function*. It was introduced in Blanchard and Roquain (2008), who called $\xi(r)$ shape function, because they studied only the case of linear thresholds under independence and positive dependence to control the FDR in that paper.

Proposition 2.3 *Let $p_1, \ldots, p_m$ be the p-values computed for the m null hypotheses satisfying* (1). *Non-negative weights $w_1, \ldots, w_m$ are given. Let $\xi(r)$ be as described in* (8). *Then the step-up procedure with threshold sequence $t_r = \xi(s(r))\alpha/m$ applied to $q_1 = p_1/w_1, \ldots, q_m = p_m/w_m$ ($q_i = 1$ if $w_i = 0$), satisfies*

$$\mathrm{sFDR} \leq \alpha/m \sum_{i=1}^{m} w_i \, .$$

Proof By case (iii) of the Lemma 3.2 of Blanchard and Roquain (2008) with $U = p_i$, $V = s(R)$ and $c = w_i\alpha/m$ the result follows. □

Blanchard and Roquain (2008) proposed different reshaping functions for the FDR case and commented on their use. For example, the reshaping function

$$\xi(r) = r/(1 + 1/2 + \cdots + 1/m) \tag{9}$$

is obtained for the probability measure ν that puts point mass proportional to $1/r$ on $r = 1, \ldots, m$, and it corresponds to the correction that Benjamini and Yekutieli (2001) applied to the linear thresholds (the BH procedure) in order to guarantee the control of the FDR under general dependence. This correction is rather conservative and may result in fewer rejections than the Bonferroni procedure. With the same reshaping function applied to a non-decreasing shape function s, we obtain the threshold sequence

$$t_r = s(r)\alpha/(m(1 + 1/2 + \cdots + 1/m)) \, , \tag{10}$$

which guarantees, with a step-up procedure, the control of the sFDR under general dependence (Proposition 2.3). For more discussion on reshaping functions, we refer to Blanchard and Roquain (2008).

3 Discussion

3.1 Relation with existing error rates

The sFDR is an extension of the FDR, based on new error rates and new multiple testing procedures. We can also define error rates based on tail probabilities such as the scaled false exceedence rate (sFER) with scale function s and rate q that we define by

$$s\mathrm{FER}_q = \mathbb{P}\left(\frac{V}{s(R)} > q\right) .$$

By Markov's inequality

$$\mathbb{E}\left(\frac{V}{q\,s(R)}\right) \leq \alpha \ \Rightarrow s\text{FER}_q = \mathbb{P}\left(\frac{V}{s(R)} > q\right) \leq \alpha,$$

from which it follows that the step-up procedure with threshold sequence

$$t_r = qs\,(r)\,\alpha/m, \quad r = 1, \ldots, m\,, \tag{11}$$

provides conservative control of this error rate, under independence (Proposition 2.1) and positive dependence (Proposition 2.2). Under general dependency structure, we replace $qs(R)$ by $\xi(qs(R))$ (see Proposition 2.3). Based on the Lehmann and Romano (2005) heuristic, a tighter control of the sFER can be obtained by step-down procedures, also under different assumptions (Meskaldji et al. 2011).

An interesting choice is $q = 0.5$, in which case the control of the $s\text{FER}_q$ guarantees the control of the median of $V/s(R)$ and we note that in order to achieve control, we simply have to halve α. Consequently, if we wish to control $median(FDP)$ at level α, we can use the BH procedure with $\alpha/2$.

From the sFDR and the sFER, one can recover most error rates that have been introduced in the literature and can derive the corresponding control procedures under different assumptions, in the similar way as presented in the introduction. For more examples, refer to (Meskaldji et al. 2011).

Error rates related to ours were introduced in van der Laan et al. (2004) and described in Dudoit and van der Laan (2008, p. 238 and ch. 6, 7). These references consider transformations that involve both V and R, while we concentrate on transforming R alone as it is more tractable and easier to control.

Other studies have proposed threshold sequences different from the horizontal or the linear, among them Genovese and Wasserman (2002, p. 508 and p. 513), where it is shown that asymptotically the linear one is optimal in some sense; Finner et al. (2009), who derive an alternative threshold sequence based on asymptotic arguments; and Roquain and Villers (2011), who investigate the operating characteristics of the FDR under an increasing threshold sequence. But none of these papers makes the connection to a generalization of the error rate.

3.2 Approximation under independence

Under the random mixture effect model introduced in Efron et al. (2001) (in which, p-values are independent), when applying a procedure that exerts FDR control, we can expect to have (via heuristic arguments) $V \approx \frac{m_0}{m}\alpha R$ for $m >> 1$ (Genovese and Wasserman 2002; Roquain and Villers 2011), which means that the number of false positives increases as a fraction of the total number of effects found.

Similarly, under the same model, if a step-up procedure is applied with the threshold sequence $t_r = s(r)\alpha/m$, the number of false positives can be approximated by

$$V \approx \frac{m_0}{m}\alpha s(R) \tag{12}$$

when the number of tests m is large. This approximation indicates how the s function acts as a scaling moderating the control of the false positives for different possible values of the number of rejections R. This can be further highlighted by

$$\frac{V}{s(R)} = \frac{V}{R} \cdot \frac{R}{s(R)}, \tag{13}$$

which shows that control of the sFDP is equivalent to control of the FDP times a positive multiplier $\frac{R}{s(R)}$ that depends on R and that could be greater or less than 1 depending on the level of conservativeness that the researcher desires. Furthermore, when the scale function s is a constant, the sFDP depends only on V and the dependence on the number of rejections R is suppressed.

Based on (13), the threshold can also be interpreted as a weight, albeit one that applies to the rank of the hypothesis as determined by its p-value. If we re-write $p_{(r)} \leq s(r)\alpha/m$ as $p_{(r)}r/s(r) \leq r\alpha/m$, we see that the method acts as a weighted BH procedure with weight $s(r)/r$ for $p_{(r)}$, $r = 1, \ldots, m$. Since these weights depend on the p-values, the condition $\sum_r^m s(r)/r = m$ does not guarantee the control of the FDR.

3.3 The choice of the threshold sequence

The choice of the error rate control criterion should depend on the particular research question in a practical data problem (Pigeot 2000). In our approach, the function s serves two purposes. It defines the threshold sequence and the error rate. The freedom offered by using any s function is practical and useful. Instead of having to choose between specific criteria such as FDR or FWER, $\mathbb{E}(V)$ and FER, it gives users the opportunity to choose on a very fine scale at one end, tests that merely provide a screening of a large number of potential effects, and at the other end, tests that exert a strict control on the number of false discoveries, at the cost of difficulties in finding moderately large true effects. This can be achieved by specifying a scale function s, which allows the user to moderate the proportion between the false positives and the number of rejections, and build the corresponding control procedure under different assumptions.

In the following, we restrict the discussion to the case of concave shape functions, which was the motivation for studying other thresholds in our original work (Meskaldji et al. 2011). The linear thresholds $t_r = r\alpha/m$ are ideal in situations where the number of tests is not very large and the number of true effects is quite small and the effects are just barely detectable. In other situations, the number R of detected effects can be large or moderately large and the linear thresholds will then find a sizable number of false positives and lose their effectiveness. The concave threshold is meant to capture the p-values belonging to true effects from above. In Fig. 1, such p-values cluster near the bottom left corner, that is, they are small and have low rank. This is the situation we have in mind, even though this will not happen if the true effects are not large enough. If we move to the right along the curve of the ordered p-values, the ones belonging to true effects will become rare and the p-values sequence cuts across the threshold.

Concretely, consider a case where the number of tests is relatively large and we apply the BH procedure with a typical $\alpha = 0.05$. In some cases, the number of rejections suits us and the approximate expected number of false positives is relatively small. In other cases, the number of rejections may also be very small, indicating a small number of alternatives or weak effect size. In this case we would have liked to use a larger α, 0.2 for example. But α cannot be chosen based on the observed p-values without losing control with non-adaptive methods. Inversely, the number of rejections can be very large, more than 1000 for example. For $\alpha = 0.05$ we expect to have a number of false positives of the order of 50 (in the independent case), which can be problematic in some applications. In this case, one can eliminate the 50 largest p-values or decrease α to 0.01 for example. Again, these naive approaches do not guarantee any control with non-adaptive methods because the choice of α is based on the p-values. We may want to have a procedure that has the same behavior as the FDR control when the expected number of false positives is reasonable, and which anticipates the case of few or weak effects, behaving as a FDR control with a larger α, and at the same time anticipating the case of strong alternatives, behaving like a FDR control with a smaller α. This flexibility can be provided by a concave s function.

For example, let us consider a concave function s and the associated threshold sequence $s(r)\alpha/m, r = 1, \ldots, m$, that intersect with the linear threshold sequence $r\alpha/m, r = 1, \ldots, m$, at $k > 0$, that is $s(k) = k$. The point of intersection between the concave function and the linear function has to be selected by the user and it depends on the problem at hand. Let R, R_s be the number of rejections when performing a step-up procedure with the linear and concave threshold sequences, respectively. Similarly, let V, V_s be the corresponding number of false positives, respectively. If the last crossing of the p-values happens before k, that is, if $R < k$, then $R_s \geq R$. In this case, the expected number of false positives of the concave threshold is larger but remains $\mathbb{E}(V_s) < k\alpha$. On the other hand, if the last crossing occurs after the index k, more false negatives will be tolerated and $R_s \leq R$ but the number of false positives will be dramatically reduced if $s(R) << R$ for $R >> k$. It is true that the first situation can be obtained by using the identity shape function (linear threshold sequence) with a larger slope (a larger α), while the second situation can be obtained with a smaller slope (smaller α). However, the two conditions can never be satisfied at the same time with a linear threshold sequence or a horizontal one. Because the choice of α cannot be made a posteriori (at the risk of losing control at least in the non-adaptive case), and in the absence of any a priori concerning the real parameters of the problem, the concave choice seems to be the best compromise to gain more power in the case of few and weak alternatives (small number of rejections with the linear threshold), while exerting a stricter control on the false positives in the case of lot and strong alternatives (inducing a large number of rejections with the linear threshold). The use of a convex shape function will result in an inverse behavior and could be suitable in some applications.

In Table 3, we present simulations of thresholds with concave shape functions of the form $s(r) = C\, r^{\gamma}$ (where C is a constant) that intersect with the linear threshold sequence at the point of coordinates $(k, k\,\alpha/m)$. We also considered the horizontal thresholds that intersect with the concave and the linear threshold sequences at the same point $(k, k\,\alpha/m)$. In all cases, the parameter γ was set to 0.5 and the parameter

Table 3 Simulation of the independent mixture effect for different values of the parameters m, m_1 and Δ

m	m_1	Δ	V_s	S_s	V_h	S_h	V_l	S_l
1000	10	1	0.241	0.006	0.993	0.016	0.054	0.003
1000	10	2	0.254	0.029	0.987	0.050	0.056	0.017
1000	10	3	0.269	0.111	1.004	0.145	0.058	0.078
1000	50	1	0.245	0.153	0.943	0.286	0.053	0.076
1000	50	2	0.297	0.846	0.932	1.205	0.081	0.556
1000	50	3	0.497	4.042	0.958	4.592	0.230	3.424
1000	100	1	0.281	0.631	0.897	1.074	0.069	0.331
1000	100	2	0.452	3.868	0.877	4.748	0.202	2.933
1000	100	3	0.856	16.258	0.900	16.447	0.815	16.032
1000	200	1	0.368	2.813	0.805	3.976	0.127	1.740
1000	200	2	0.762	16.937	0.791	17.172	0.750	16.658
1000	200	3	1.367	55.855	0.806	50.960	2.657	62.259
10000	50	1	0.429	0.013	2.502	0.032	0.051	0.003
10000	50	2	0.437	0.039	2.485	0.067	0.049	0.018
10000	50	3	0.516	0.244	2.615	0.353	0.077	0.166
10000	500	1	0.483	0.735	2.401	1.470	0.064	0.302
10000	500	2	0.950	5.953	2.400	7.943	0.260	3.954
10000	500	3	2.012	34.275	2.344	35.264	1.699	32.931
10000	1000	1	0.757	3.561	2.295	5.803	0.174	1.766
10000	1000	2	1.795	29.670	2.236	31.610	1.321	27.012
10000	1000	3	3.888	141.185	2.272	127.496	7.630	160.189
10000	2000	1	1.309	18.538	2.015	22.332	0.656	13.743
10000	2000	2	3.331	133.263	1.950	114.783	6.904	164.755
10000	2000	3	6.455	488.889	2.036	393.926	25.976	624.406

The observations are distributed as $\mathcal{N}(0, 1)$ or $\mathcal{N}(\Delta, 1)$, respectively. For the concave case, the parameter $k = 20$ when $m = 1000$ and $k = 50$ when $m = 10000$. The parameter α was set to 0.05. (V_s, S_s), (V_h, S_h) and (V_l, S_l) represent the average of false positives and true positives with the concave, horizontal and linear cases, respectively. The values are obtained on the basis of 1000 simulations in each case

α was set to 0.05. The observations are independent and distributed as $\mathcal{N}(0, 1)$ or $\mathcal{N}(\Delta, 1)$ for true null and false null hypotheses, respectively.

As suggested by the reviewers, we will not concentrate on a particular gain function and will simply report the average values of the number of false positives and the number of true positives in each case. (V_s, S_s), (V_h, S_h) and (V_l, S_l) represent the average of false positives and true positives with the concave, horizontal and linear cases, respectively. Note that for large Δ and m_1, the average of S_s is larger than the average of S_h while keeping a low average value of V_s compared to V_l. For small Δ and m_1, the concave threshold makes detections more than the linear, while keeping the false positives under control, even less than the horizontal.

3.4 The truncated false discovery rate

In applied statistics, interpretation is of high importance, and the control of the sFDR may be difficult to interpret for many users. We propose a particular family of concave functions that offers a simple choice that can be made in practice, namely threshold sequences that grow linearly in r for small values ($r \leq k$ for a predefined $k \geq 1$), and reduces its slope for larger values ($r \geq k$). The prototype is the trimmed threshold with truncated shape function $s^{trunc}(r) = \min(k, r)$ for $1 \leq k \leq m$.

Under dependence and positive dependence, the corresponding step-up procedure with threshold sequence $t_r^{\text{trunc}} = s^{trunc}(r)\alpha/m$ controls the FDR to be less than α and controls $\mathbb{E}(V)$ to be less than $k\alpha$ as the Hommel and Hoffmann (1987) procedure guarantees. We might call this type of control, that is, the control of $\mathbb{E}(V/s^{trunc}(R))$ the truncated FDR. This has a very simple interpretation and is an appropriate choice, if the number m of hypotheses being tested is very large. Applications include functional magnetic resonance images in the neuroscience as well as many genomic or other "omic" studies (Meskaldji et al. 2015, 2013).

For general dependence, we can use the same reshaping function $\xi(r) = r/(1 + 1/2 + \cdots + 1/m)$ as in Benjamini and Yekutieli (2001). However, if we instead take ν as the probability measure with point mass proportional to $1/1,\ 1/2, \ldots$ with support $\{1, 2, \ldots, k\}$, the reshaping function truncates, because it is equal to

$$\xi(r) = r\left(\sum_{i=1}^{k} \frac{1}{i}\right)^{-1} \quad \text{for } r \leq k \text{ and } \xi(r) = \xi(k) \text{ for } r > k\,.$$

The corresponding threshold sequence is

$$s(r)^{\text{trunc}}\alpha \Bigg/ \left(m\sum_{i=1}^{k} \frac{1}{i}\right),$$

which exceeds (10) for $r \leq \lfloor k \sum_{i=1}^{m} \frac{1}{i} / \sum_{i=1}^{k} \frac{1}{i} \rfloor$. This test thus increases the power in some situations even though it exerts a provably stricter control (see Fig. 2). A particular case is $k = 1$, where $\xi(r) \equiv 1$, and the threshold sequence reduces to $\alpha/m, r = 1, \ldots, m$, that is, the Bonferroni procedure.

Note that the truncated FDR could be written as

$$\mathbb{E}\left(\frac{V}{s^{\text{trunc}}(R)}\right) = \mathbb{E}\left(\frac{V}{\min(k, R)}\right) \geq \max\left[\mathbb{E}\left(\frac{V}{R}\right), \mathbb{E}\left(\frac{V}{k}\right)\right]. \tag{14}$$

A similar criterion (with smooth transition) is obtained by minimizing the expected number of false positives while simultaneously controlling the false discovery rate. This leads to a family of mixed error rates (MER) indexed by $0 \leq \epsilon \leq 1$,

$$\text{MER}_{\epsilon,k} = \epsilon\mathbb{E}(V/k) + (1 - \epsilon)\mathbb{E}(V/R)\,,$$

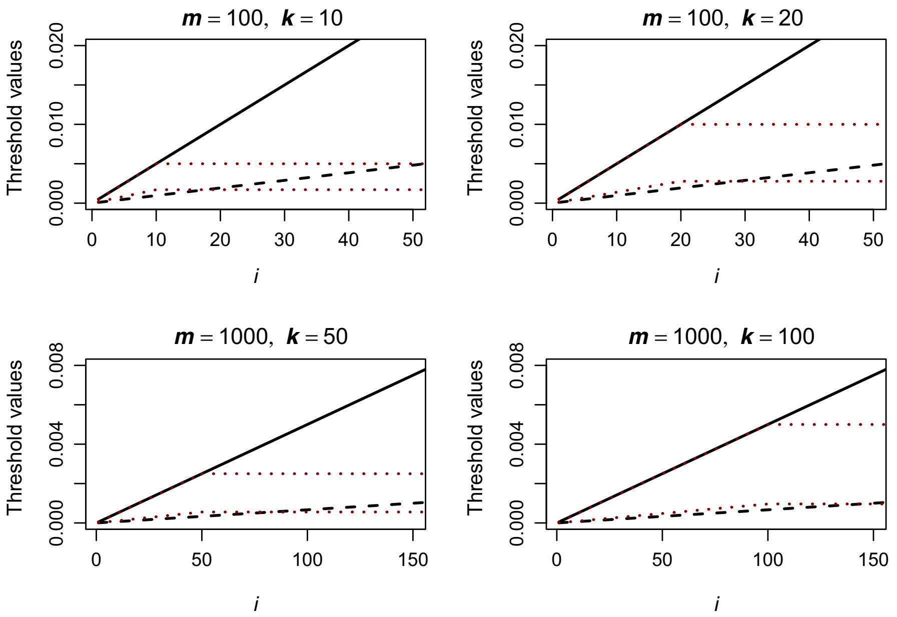

Fig. 2 Examples of threshold sequences for different values of m and k. The BH procedure (solid line) and the truncated linear procedure (dash-dotted line), both for independent and positively dependent tests. The modified BH thresholds as described in Benjamini and Yekutieli (2001) (dashed line) and the modified truncated linear thresholds as described in the Results section (dotted line), both for the general dependence case

where ϵ tunes the transition between the two controls. To derive a step-up procedure that controls the MER, we use Propositions 2.1 and 2.2. We have

$$\mathrm{MER}_{\epsilon,k} = \mathbb{E}\left[V \left(\frac{R}{\epsilon R/k + (1-\epsilon)} \right)^{-1} \right],$$

which we recognize as a sFDR with scale function

$$s^{\epsilon,k}(R) = \frac{R}{\epsilon R/k + (1-\epsilon)} .$$

Control at level α, when m independent or positively dependent hypotheses are tested, is obtained by the step-up procedure with threshold sequence

$$t_r^{\epsilon,k} = s^{\epsilon,k}(r)\alpha/m = \frac{r}{\epsilon r/k + (1-\epsilon)}\alpha/m .$$

4 Conclusion

We introduced a new indicator for the control of false positives, the scaled false discovery proportion sFDP, and we proposed two error rates, the sFDR and the sFER. We showed that in this framework, one can embed the classical error rates in a family of multiple testing procedures, for example, FWER and FER or $\mathbb{E}(V)$ and FDR. The freedom offered by the scaling function generalizes the existent error rates and offers the user a finer control between tests whose aim is a screening of the hypotheses and tests whose aim is the detection of true alternatives. We also proposed the corresponding multiple testing procedures to control either the sFDR or the sFER under varying assumptions. Other classes of procedures could be generalized in the same way as presented in this paper.

We discussed the case of concave scaling functions, in particular, truncated linear scaling functions, which represent an intermediate choice between FWER and FDR, and we presented simulations to show their benefit in some practical situations.

In addition to the freedom of choice of error rate, the duality that exists in

$$t_r = \xi(s(r))\alpha/m \equiv \tilde{s}(r)\alpha/m, \quad \text{for } r = 1, \ldots, m,$$

between the error rate and the threshold sequence $t_r = s_r\alpha/m$ together with the reshaping function $\xi(r)$, brings new perspectives to the field of multiple testing procedures and reveals new interesting results.

Acknowledgements The main part of this work was done while the first author was a Ph.D student in the Signal Processing Laboratory (LTS5), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland (Meskaldji 2013). The authors would like to thank Etienne Roquain, Arnold Janssen and Bradley Efron for interesting comments and suggestions. We also thank the referees for their helpful and constructive comments.

Funding This work was supported by the Swiss National Science Foundation [144467, PP00P2-146318].

References

Benjamini Y (2010) Discovering the false discovery rate. J R Stat Soc Ser B 72(4):405–416

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57(1):289–300

Benjamini Y, Hochberg Y (1997) Multiple hypotheses testing with weights. Scand J Stat 24(3):407–418

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29(4):1165–1188

Bernhard G, Klein M, Hommel G (2004) Global and multiple test procedures using ordered p-values—a review. Stat Pap 45(1):1–14

Blanchard G, Roquain E (2008) Two simple sufficient conditions for FDR control. Electron J Stat 2:963–992

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. Pub del R Ist Sup di Sci Eco e Com di Fir 8:3–62 Bonferroni adjustment for multiple statistical tests using the same data

Dudoit S, van der Laan MJ (2008) Multiple testing procedures with applications to genomics. Springer Series in Statistics. Springer, New York

Efron B, Tibshirani R, Storey J, Tusher V (2001) Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96:1151–1160

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

Genovese C, Wasserman L (2002) Operating characteristics and extensions of the false discovery rate procedure. J R Stat Soc Ser B 64:499–517

Genovese CR, Roeder K, Wasserman L (2006) False discovery control with p-value weighting. Biometrika 93(3):509–524

Heesen P, Janssen A (2015) Inequalities for the false discovery rate (fdr) under dependence. Electron J Stat 9(1):679–716

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75(4):800–802

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat Theory Appl 6(2):65–70

Hommel G, Hoffmann T (1987) Controlled uncertainty. In: Bauer P, Hommel G, Sonnemann E (eds) Multiple hypotheses testing. Springer, Heidelberg, pp 154–161

Lehmann EL, Romano JP (2005) Generalizations of the familywise error rate. Ann Stat 33(3):1138–1154

Meskaldji D-E (2013) Multiple comparison procedures for large correlated data with application to brain connectivity analysis. Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne

Meskaldji D-E, Fischi-Gomez E, Griffa A, Hagmann P, Morgenthaler S, Thiran J-P (2013) Comparing connectomes across subjects and populations at different scales. Neuroimage 80:416–425 Mapping the Connectome

Meskaldji D-E, Thiran J-P, Morgenthaler S (2011) A comprehensive error rate for multiple testing. ArXiv e-prints arXiv:1112.4519

Meskaldji D-E, Vasung L, Romascano D, Thiran J-P, Hagmann P, Morgenthaler S, Ville DVD (2015) Improved statistical evaluation of group differences in connectomes by screening-filtering strategy with application to study maturation of brain connections between childhood and adolescence. Neuroimage 108:251–264

Pigeot I (2000) Basic concepts of multiple tests—a survey. Stat Pap 41(1):3–36

Roquain E (2015) Contributions to multiple testing theory for high-dimensional data. Université Pierre et Marie Curie, Paris

Roquain E, Villers F (2011) Exact calculations for false discovery proportion with application to least favorable configurations. Ann Stat 39:584–612

Sarkar SK (1998) Some probability inequalities for ordered MTP2 random variables: a proof of the simes conjecture. Ann Stat 26(2):494–504

Schweder T, Spjøtvoll E (1982) Plots of p-values to evaluate many tests simultaneously. Biometrika 69(3):493–502

Simes RJ (1986) An improved Bonferroni procedure for multiple tests of significance. Biometrika 73(3):751–754

van der Laan MJ, Dudoit S, Pollard KS (2004) Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. Stat Appl Genet Mol Biol 3:1–25 Art. 15 (electronic)