

Type I error rate control for testing many hypotheses: a survey with proofs

Titre: Une revue du contrôle de l'erreur de type I en test multiple

Etienne Roquain¹

Abstract: This paper presents a survey on some recent advances for the type I error rate control in multiple testing methodology. We consider the problem of controlling the k -family-wise error rate (kFWER, probability to make k false discoveries or more) and the false discovery proportion (FDP, proportion of false discoveries among the discoveries). The FDP is controlled either via its expectation, which is the so-called false discovery rate (FDR), or via its upper-tail distribution function. We aim at deriving general and unified results together with concise and simple mathematical proofs. Furthermore, while this paper is mainly meant to be a survey paper, some new contributions for controlling the kFWER and the upper-tail distribution function of the FDP are provided. In particular, we derive a new procedure based on the quantiles of the binomial distribution that controls the FDP under independence.

Résumé : Ce travail présente une revue des récents travaux du contrôle de l'erreur de type I en test multiple. On considère le problème du contrôle du " k -family-wise error rate" (kFWER, probabilité d'effectuer au moins k fausses découvertes) et du "false discovery proportion" (FDP, proportion de fausses découvertes parmi les découvertes). Le FDP est contrôlé soit via son espérance (correspondant au fameux "false discovery rate") soit via sa queue de distribution. Nous recherchons à obtenir à la fois des résultats unifiés et des preuves mathématiques simples et concises. De plus, nous proposons de nouvelles contributions méthodologiques pour contrôler le kFWER et la queue de distribution du FDP. En particulier, nous introduisons une nouvelle procédure qui contrôle le FDP sous indépendance et qui est basée sur les quantiles de la loi binomiale.

Keywords: multiple testing, type I error rate, false discovery proportion, family-wise error, step-up, step-down, positive dependence

Mots-clés : test multiple, erreur de type I, taux de fausses découvertes, probabilité de fausses découvertes, dépendance positive

AMS 2000 subject classifications: 62J15, 62G10

¹ UPMC University of Paris 6, LPMA.
E-mail: etienne.roquain@upmc.fr

1. Introduction

The problem of testing several null hypotheses has a long history in the statistics literature. With the high-resolution techniques introduced in the recent years, it has known a renewed attention in many application fields where one aims to find significant features among several thousands (or millions) of candidates. Classical examples are microarray analysis [58, 17, 19, 20], neuro-imaging analysis [4, 42] and source detection [38]. For illustration, we detail below the case of microarray data analysis.

1.1. Multiple testing in microarray data

In a typical microarray experiment, the level expressions of a set of genes are measured under two different experimental conditions and we aim at finding the genes that are differentially expressed between the two conditions. For instance, when the genes come from tumor cells in the first experimental condition, while they come from healthy cells in the second, the differentially expressed genes may be involved in the development of this tumor and thus are genes of special interest. Several techniques exist to perform a statistical test for a single gene, e.g. based on a distributional assumption or on permutations between the two group labels. However, the number of genes m can be large (for instance several thousands), so that non-differentially expressed genes can have a high score of significance by chance. In that context, applying the naive, non-corrected procedure (level α for each gene) is unsuitable because it is likely to select (or “discover”) a lot of non-differentially expressed genes (usually called “false discoveries”). For instance, if the $m = 10,000$ genes are not differentially expressed (no signal) and $\alpha = 0.1$, the non-corrected procedure makes on average $m\alpha = 1,000$ discoveries which are all false discoveries. In a more favorable situation where there are only $m_0 = 5,000$ non-differentially expressed genes among the $m = 10,000$ initial genes (50% of signal), the non-corrected procedure selects some genes, say r genes, for which the expected number of errors is $m_0\alpha = 500$. Since the number of discoveries r is not designed to be much larger than the number of false discoveries $m_0\alpha$, the final list of discovered genes is likely to contain an unacceptable part of errors. A multiple testing procedure aims at correcting *a priori* the level of the single tests in order to obtain a list of selected genes for which the “quantity” of false discoveries is below a nominal level α . The “quantity” of false discoveries is measured by using *global type I error rates*, as for instance the probability to make at least k errors among the discoveries (k -family-wise error rate, k -FWER) or the expected proportion of errors among the discoveries (false discovery rate, FDR). Finding procedures that control type I error rates is challenging and is what we called here the “multiple testing issue”. Furthermore, a feature that increases the complexity of this issue is the presence of dependencies between the single tests.

Note that the multiple testing issue can be met in microarray analysis under other forms, as for instance when we search co-expressed genes or genes associated with clinical covariates or outcomes, see Section 1.2 of [17].

1.2. Examples of multiple testing settings

Example 1.1 (Two-sample multiple t -tests). The problem of finding differentially expressed genes in the above microarray example can be formalized as a particular case of a general two-sample

multiple testing problem. Let us observe a couple of two independent samples

$$X = (X^1, \dots, X^n) = (Y^1, \dots, Y^{n_1}, Z^1, \dots, Z^{n_2}) \in \mathbb{R}^{m \times n},$$

where (Y^1, \dots, Y^{n_1}) is a family of n_1 i.i.d. copies of a random vector Y in \mathbb{R}^m and (Z^1, \dots, Z^{n_2}) is a family of n_2 i.i.d. copies of a random vector Z in \mathbb{R}^m (with $n_1 + n_2 = n$). In the context of microarray data, Y_i^j (resp. Z_i^j), $1 \leq i \leq m$, corresponds to the expression level measure of the i -th gene for the j -th individual of the first (resp. second) experimental condition. Typically, the sample size is much smaller than the number of tests, that is, $n \ll m$. Let the distribution P of the observation X belong to a statistical model given by a distribution set \mathcal{P} . Assume that \mathcal{P} is such that X is an integrable random vector and let $\mu_{i,1}(P) = \mathbb{E}Y_i$ and $\mu_{i,2}(P) = \mathbb{E}Z_i$, for any $i \in \{1, \dots, m\}$. The aim is to decide for all i whether P belongs to the set $\Theta_{0,i} = \{P \in \mathcal{P} : \mu_{i,1}(P) = \mu_{i,2}(P)\}$ or not, that is, we aim at testing the hypothesis

$$H_{0,i} : \text{"}\mu_{i,1}(P) = \mu_{i,2}(P)\text{" against } H_{1,i} : \text{"}\mu_{i,1}(P) \neq \mu_{i,2}(P)\text{"},$$

simultaneously for all $i \in \{1, \dots, m\}$. Given P , the null hypothesis $H_{0,i}$ (sometimes called the "null" for short) is said to be true (for P) if $P \in \Theta_{0,i}$, that is, if P satisfies $H_{0,i}$. It is said false (for P) otherwise. The index set corresponding to true nulls is denoted by $\mathcal{H}_0(P) = \{1 \leq i \leq m : \mu_{i,1}(P) = \mu_{i,2}(P)\}$. Its complement in $\mathcal{H} = \{1, \dots, m\}$ is denoted by $\mathcal{H}_1(P)$. In the microarray context, $\mathcal{H}_1(P) = \{1 \leq i \leq m : \mu_{i,1}(P) \neq \mu_{i,2}(P)\}$ is thus the index set corresponding to differentially expressed genes. The aim of a multiple testing procedure is thus to recover the (unobservable) set $\mathcal{H}_1(P)$ given the observation X . A multiple testing procedure is commonly based on individual test statistics, by rejecting the null hypotheses with a "large" test statistic. Here, the individual test statistic can be the (two-sided) two-sample t-statistic $S_i(X) \propto |\bar{Y}_i - \bar{Z}_i|$, rescaled by the so-called "pooled" standard deviation. To provide a uniform normalization for all tests, it is convenient to transform the $S_i(X)$ into the p -value

$$p_i(X) = \sup_{P \in \Theta_{0,i}} T_{P,i}(S_i(X)), \quad (1)$$

where $T_{P,i}(s) = \mathbb{P}_{X \sim P}(S_i(X) \geq s)$ is the upper-tail distribution function of $S_i(X)$ for $X \sim P \in \Theta_{0,i}$. Classically, assuming that Y_i and Z_i are Gaussian variables with the same variance, we have for any $P \in \Theta_{0,i}$, $T_{P,i}(s) = 2\mathbb{P}(Z \geq s)$, where Z follows a Student distribution with $n - 2$ degrees of freedom. In that case, each p -value $p_i(X)$ has the property to be uniformly distributed on $(0, 1)$ when the corresponding null hypothesis $H_{0,i}$ is true. Without making this Gaussian assumption, p -values can still be built, as we discuss in Remark 1.3 below. Let us finally note that since the $T_{P,i}$ are decreasing, a multiple testing procedure should reject nulls with a "small" p -value.

Example 1.2 (One-sided testing on the mean of a Gaussian vector). To give a further illustrating example, we consider the very convenient mathematical framework for multiple testing where we observe a Gaussian vector $X = (X_i)_{1 \leq i \leq m} \sim P$, having an unknown mean $\mu(P) = (\mu_i(P))_{1 \leq i \leq m} \in \mathbb{R}^m$ and a $m \times m$ covariance matrix $\Sigma(P)$ with diagonal entries equal to 1. Let us consider the problem of testing

$$H_{0,i} : \text{"}\mu_i(P) \leq 0\text{" against } H_{1,i} : \text{"}\mu_i(P) > 0\text{"},$$

simultaneously for all $i \in \{1, \dots, m\}$. We can define the p -values $p_i = \bar{\Phi}(X_i)$, where $\bar{\Phi}(x) = \mathbb{P}(Z \geq x)$ for $Z \sim \mathcal{N}(0, 1)$. Any p -value satisfies the following stochastic domination under the null: if

$\mu_i(P) \leq 0$, we have for all $u \in [0, 1]$,

$$\mathbb{P}(p_i(X) \leq u) \leq \mathbb{P}(\bar{\Phi}(X_i - \mu_i(P)) \leq u) = u.$$

Additionally, more or less restrictive assumptions on $\Sigma(P)$ can be considered to model different types of dependency of the corresponding p -values. For instance, we can assume that $\Sigma(P)$ has only non-negative entries, that the non-diagonal entries of $\Sigma(P)$ are equal (equi-correlation) or that $\Sigma(P)$ is diagonal. Finally, the value of the alternative means can be used for modeling the “strength of the signal”. For instance, to model that the sample size available for each test is n , we can set $\mu_i(P) = \tau\sqrt{n}$ for each $\mu_i(P) > 0$, where $\tau > 0$ is some additional parameter.

Remark 1.3 (General construction of p -values). In broad generality, when testing the nulls $\Theta_{0,i}$ by rejecting for “large” values of a test statistic $S_i(X)$, we can always define the associated p -values by using (1). It is well known that these p -values are always stochastically lower-bounded by a uniform variable under the null, that is, $\forall i \in \mathcal{H}_0(P), \forall u \in [0, 1], \mathbb{P}(p_i(X) \leq u) \leq u$. This property always holds, even when $S_i(X)$ has a discrete distribution. For completeness, we provide this result with a proof in Appendix A. However, the calculation of the p -values (1) is not always possible, because it requires the knowledge of the distribution of the test statistics under the null, which often relies on strong distributional assumptions on the data. Fortunately, in some situations, the p -values (1) can be approximated by using a randomization technique. The resulting p -values can be shown to enjoy the same stochastic dominance as above (see, e.g., [44] for a recent reference). For instance, in the two-sample testing problem, permutations of the group labels can be used, which corresponds to use permutation tests (the latter can be traced back to Fisher [25]).

1.3. General multiple testing setting

In this section, we provide the abstract framework in which multiple testing theory can be investigated in broad generality.

Let us consider a statistical model, defined by a measurable space $(\mathcal{X}, \mathfrak{X})$ endowed with a subset \mathcal{P} of distributions on $(\mathcal{X}, \mathfrak{X})$. Let X denote the observation of the model, with distribution $P \in \mathcal{P}$. Consider a family $(\Theta_{0,i})_{1 \leq i \leq m}$ of $m \geq 2$ subsets of \mathcal{P} . Based on X , we aim at testing the null hypotheses $H_{0,i} : “P \in \Theta_{0,i}”$ against the alternative $H_{1,i} : “P \in \Theta_{0,i}^c”$ simultaneously for all $i \in \{1, \dots, m\}$. For any $P \in \mathcal{P}$, let $\mathcal{H}_0(P) = \{1 \leq i \leq m : P \in \Theta_{0,i}\}$ be the set of the indexes i for which P satisfies $H_{0,i}$, that is, the indexes corresponding to true null hypotheses. Its cardinality $|\mathcal{H}_0(P)|$ is denoted by $m_0(P)$. Similarly, the set $\{1, \dots, m\}$ is sometimes denoted by \mathcal{H} . The set of the false null hypotheses is denoted by $\mathcal{H}_1(P) = \mathcal{H} \setminus \mathcal{H}_0(P)$. The goal is to recover the set $\mathcal{H}_1(P)$ based on X , that is, to find the null hypotheses that are true/false based on the knowledge of X . Obviously, the distribution P of X is unknown, and thus so is $\mathcal{H}_1(P)$.

The standard multiple testing setting includes the knowledge of p -values $(p_i(X))_{1 \leq i \leq m}$ satisfying

$$\forall P \in \mathcal{P}, \forall i \in \mathcal{H}_0(P), \forall u \in [0, 1], \mathbb{P}(p_i(X) \leq u) \leq u. \quad (2)$$

As a consequence, for each $i \in \{1, \dots, m\}$, rejecting $H_{0,i}$ whenever $p_i(X) \leq \alpha$ defines a test of level α . As we have discussed in the previous section, property (2) can be fulfilled in many situations. Also, in some cases, (2) holds with equality, that is, the $p_i(X)$ are exactly distributed like a uniform variable in $(0, 1)$ when $H_{0,i}$ is true.

1.4. Multiple testing procedures

In the remainder of the paper, we use the observation X only through the p -value family $\mathbf{p}(X) = \{p_i(X), 1 \leq i \leq m\}$. Therefore, for short, we often drop the dependence in X in the notation and define all quantities as functions of $\mathbf{p} = \{p_i, 1 \leq i \leq m\} \in [0, 1]^m$. However, one should keep in mind that the underlying distribution P (the distribution of interest on which the tests are performed) is the distribution of X and not the one of \mathbf{p} .

A *multiple testing procedure* is defined as a set-valued function

$$R : q = (q_i)_{1 \leq i \leq m} \in [0, 1]^m \mapsto R(q) \subset \{1, \dots, m\},$$

taking as input an element of $[0, 1]^m$ and returning a subset of $\{1, \dots, m\}$. For such a general procedure R , we add the technical assumption that for each $i \in \{1, \dots, m\}$, the mapping $x \in \mathcal{X} \mapsto \mathbf{1}\{i \in R(\mathbf{p}(x))\}$ is measurable. The indexes selected by $R(\mathbf{p})$ correspond to the rejected null hypotheses, that is, $i \in R(\mathbf{p}) \Leftrightarrow "H_{0,i} \text{ is rejected by the procedure } R(\mathbf{p})"$. Thus, for each p -value family \mathbf{p} , there are 2^m possible outcomes for $R(\mathbf{p})$. Nevertheless, according to the stochastic dominance property (2) of the p -values, a natural rejection region for each $H_{0,i}$ is of the form $p_i \leq t_i$, for some $t_i \in [0, 1]$. In this paper, we mainly focus on the case where the threshold is the same for all p -values. The corresponding procedures, called *thresholding based procedures*, are of the form $R(\mathbf{p}) = \{1 \leq i \leq m : p_i \leq t(\mathbf{p})\}$, where the threshold $t(\cdot) \in [0, 1]$ can depend on the data.

Example 1.4 (Bonferroni procedure). The Bonferroni procedure (of level $\alpha \in (0, 1)$) rejects the hypotheses with a p -value smaller than α/m . Hence, with our notation, it corresponds to the procedure $R(\mathbf{p}) = \{1 \leq i \leq m : p_i \leq \alpha/m\}$.

1.5. Type I error rates

To evaluate the quality of a multiple testing procedure, various error rates have been proposed in the literature. According to the Neyman-Pearson approach, type I error rates are of primary interest. These rates evaluate the importance of the null hypotheses wrongly rejected, that is, of the elements of the set $R(\mathbf{p}) \cap \mathcal{H}_0(P)$. Nowadays, the most widely used type I error rates are the following. For a given procedure R ,

- the *k-family-wise error rate* (k -FWER) (see e.g. [32, 44, 36]) is defined as the probability that the procedure R makes at least k false rejections: for all $P \in \mathcal{P}$,

$$k\text{-FWER}(R, P) = \mathbb{P}(|R(\mathbf{p}) \cap \mathcal{H}_0(P)| \geq k), \quad (3)$$

where $k \in \{1, \dots, m\}$ is a pre-specified parameter. In the particular case where $k = 1$, this rate is simply called the *family-wise error rate* and is denoted by $\text{FWER}(R, P)$.

- the *false discovery proportion* (FDP) (see e.g. [53, 5, 36]) is defined as the proportion of errors in the set of the rejected hypotheses: for all $P \in \mathcal{P}$,

$$\text{FDP}(R(\mathbf{p}), P) = \frac{|R(\mathbf{p}) \cap \mathcal{H}_0(P)|}{|R(\mathbf{p})| \vee 1}, \quad (4)$$

where $|R(\mathbf{p})| \vee 1$ denotes the maximum of $|R(\mathbf{p})|$ and 1. The role of the term “ $\vee 1$ ” in the denominator is to prevent from dividing by zero when R makes no rejection. Since the FDP

is a random variable, it does not define an error rate. However, the following error rates can be derived from the FDP. First, the γ -upper-tail distribution of the FDP, defined as the probability that the FDP exceeds a given γ , that is, for all $P \in \mathcal{P}$,

$$\mathbb{P}(\text{FDP}(R(\mathbf{p}), P) > \gamma), \quad (5)$$

where $\gamma \in (0, 1)$ is a pre-specified parameter. Second, the false discovery rate (FDR) [5], defined as the expectation of the FDP: for all $P \in \mathcal{P}$,

$$\text{FDR}(R, P) = \mathbb{E}[\text{FDP}(R(\mathbf{p}), P)] = \mathbb{E}\left[\frac{|R(\mathbf{p}) \cap \mathcal{H}_0(P)|}{|R(\mathbf{p})| \vee 1}\right]. \quad (6)$$

Note that the probability in (5) is upper-bounded by a nominal level $\alpha \in (0, 1)$ if and only if the $(1 - \alpha)$ -quantile of the FDP distribution is upper-bounded by γ . For instance, if the probability in (5) is upper-bounded by $\alpha = 1/2$, this means that the median of the FDP is upper-bounded by γ . With some abuse, bounding the probability in (5) is called ‘‘controlling the FDP’’ from now on.

The choice of the type I error rate depends on the context. When controlling the k -FWER, we tolerate a fixed number $(k - 1)$ of erroneous rejections. By contrast, a procedure controlling (5) tolerates a small proportion γ of errors among the final rejections (from an intuitive point of view, it chooses $k \simeq \gamma|R|$). This allows to increase the number of erroneous rejections as the number of rejections becomes large. Next, controlling the FDR has become popular because it is a simple error rate based on the FDP and because it came together with the simple Benjamini-Hochberg FDR controlling procedure [5] (some dependency structure assumptions are required, see Section 3). As a counterpart, controlling the FDR does not prevent the FDP from having large variations, so that any FDR control does not necessarily have a clear interpretation in terms of the FDP (see the related discussion in Section 6.2).

Example 1.4 (Continued). The Bonferroni procedure $R(\mathbf{p}) = \{1 \leq i \leq m : p_i \leq \alpha/m\}$ satisfies the following:

$$\mathbb{E}|R(\mathbf{p}) \cap \mathcal{H}_0(P)| = \sum_{i \in \mathcal{H}_0(P)} \mathbb{P}(p_i \leq \alpha/m) \leq \alpha m_0(P)/m \leq \alpha,$$

which means that its expected number of false discoveries is below α . Using Markov’s inequality, this implies that $R(\mathbf{p})$ makes no false discovery with probability at least $1 - \alpha$, that is, for any $P \in \mathcal{P}$, $\text{FWER}(R, P) \leq \alpha$. This is the most classical example of type I error rate control.

Remark 1.5 (Case where $\mathcal{H}_0(P) = \mathcal{H}$). For a distribution P satisfying $\mathcal{H}_0(P) = \mathcal{H}$, that is when all null hypotheses are true, the FDP reduces to $\text{FDP}(R(\mathbf{p}), P) = \mathbf{1}\{|R(\mathbf{p})| > 0\}$ and we have $\text{FWER}(R, P) = \text{FDR}(R, P) = \mathbb{P}(\text{FDP}(R(\mathbf{p}), P) > \gamma) = \mathbb{P}(|R(\mathbf{p})| > 0)$. Controlling the FWER (or equivalently the FDR) in this situation is sometimes called a ‘‘weak’’ FWER control.

Remark 1.6 (Case where all null hypotheses are equal: p -value aggregation). The general framework described in Section 1.3 includes the case where all null hypotheses are identical, that is, $\Theta_{0,i} = \Theta_0$ for all $i \in \{1, \dots, m\}$. In this situation, all p -values test the same null $H_0 : ‘‘P \in \Theta_0’’$ against some alternatives contained in Θ_0^c . For instance, in the model selection framework of [3, 18, 60], each p -value is built with respect to a specific model contained in the alternative Θ_0^c . Since we have in that case $\mathcal{H}_0(P) = \mathcal{H}$ if $P \in \Theta_0$ and $\mathcal{H}_0(P) = \emptyset$ otherwise, the three quantities

$\text{FWER}(R, P)$, $\text{FDR}(R, P)$ and $\mathbb{P}(\text{FDP}(R(\mathbf{p}), P) > \gamma)$ are equal and take the value $\mathbb{P}(|R(\mathbf{p})| > 0)$ when $P \in \Theta_0$ and 0 otherwise. As a consequence, in the case where all null hypotheses are equal, controlling the FWER, the FDR or the FDP at level α is equivalent to the problem of combining p -values to build a *single testing* for H_0 which is of level α . In particular, from a procedure R that controls the FWER at level α we can derive a single testing procedure of level α by rejecting H_0 whenever $R(\mathbf{p})$ is not empty (that is, whenever $R(\mathbf{p})$ rejects at least one hypothesis). This provides a way to aggregate p -values into one (single) test for H_0 which is ensured to be of level α . As an illustration, the FWER controlling Bonferroni procedure $R = \{1 \leq i \leq m : p_i \leq \alpha/m\}$ corresponds to the single test rejecting H_0 whenever $\min_{1 \leq i \leq m} \{p_i\} \leq \alpha/m$. The Bonferroni combination of individual tests is well known and extensively used for adaptive testing (see, e.g., [54, 3, 60]). Some other examples of p -value aggregations will be presented further on, see Remark 3.9.

1.6. Goal

Let $\alpha \in (0, 1)$ be a pre-specified nominal level (to be fixed once and for all throughout the paper). The goal is to control the type I error rates defined above at level α , for a large subset of distributions $\mathcal{P}' \subset \mathcal{P}$. That is, by taking one of the above error rate $\mathcal{E}(R, P)$, we aim at finding a procedure R such that

$$\forall P \in \mathcal{P}', \mathcal{E}(R, P) \leq \alpha, \quad (7)$$

for $\mathcal{P}' \subset \mathcal{P}$ as large as possible. Obviously, R should depend on α but we omit this in the notation for short. Similarly to the single testing case, taking $R = \emptyset$ will always ensure (7) with $\mathcal{P}' = \mathcal{P}$. This means that the type I error rate control is inseparable from the problem of maximizing the power. The probably most natural way to extend the notion of power from the single testing to the multiple testing setting is to consider the expected number of correct rejections, that is, $\mathbb{E}|\mathcal{H}_1(P) \cap R|$. Throughout the paper, we often encounter the case where two procedures R and R' satisfy $R' \subset R$ (almost surely) while they both ensure the control (7). Then, the procedure R is said *less conservative* than R' . Obviously, this implies that R is more powerful than R' . This can be the case when, e.g., R and R' are thresholding-based procedures using respective thresholds t and t' satisfying $t \geq t'$ (almost surely). As a consequence, our goal is to find a procedure R satisfying (7) with a rejection set as large as possible.

Finally, let us emphasize that, in this paper, we aim at controlling (7) for any fixed $m \geq 2$ and not only when m tends to infinity. That is, the setting is non-asymptotic in the parameter m .

1.7. Overview of the paper

The remainder of the paper is organized as follows: in Section 2, we present some general tools and concepts that are useful throughout the paper. Section 3, 4 and 5 present FDR, k -FWER and FDP controlling methodology, respectively, where we try to give a large overview of classical methods in the literature. Besides, the paper is meant to have a scholarly form, accessible to a possibly non-specialist reader. In particular, all results are given together with a proof, which we aim to be as short and meaningful as possible.

Furthermore, while this paper is mostly intended to be a review paper, some new contributions with respect to the existing multiple testing literature are given in Section 4 and 5, by extending the results of [30] for the k -FWER control and the results of [45] for the FDP control, respectively.

1.8. Quantile-binomial procedure

In section 5, we introduce a novel procedure, called the *quantile-binomial procedure* that controls the FDP under independence of the p -values. This procedure can be defined as follows;

Algorithm 1.7 (Quantile-binomial procedure). *Let for any $t \in [0, 1]$ and for any $\ell \in \{1, \dots, m\}$,*

$$q_\ell(t) = \text{the } (1 - \alpha)\text{-quantile of } \mathcal{B}(m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1, t), \quad (8)$$

where $\mathcal{B}(\cdot, \cdot)$ denotes the binomial distribution and $\lfloor \gamma(\ell - 1) \rfloor$ denotes the largest integer n such that $n \leq \gamma(\ell - 1)$. Let $p_{(1)} \leq \dots \leq p_{(m)}$ be the order statistics of the p -values. Then apply the following recursion:

- Step 1: if $q_1(p_{(1)}) > \gamma$, stop and reject no hypothesis. Otherwise, go to step 2;
- Step $\ell \in \{2, \dots, m\}$: if $q_\ell(p_{(\ell)}) > \gamma^\ell$, stop and reject the hypotheses corresponding to $p_{(1)}, \dots, p_{(\ell-1)}$. Otherwise, go to step $\ell + 1$;
- Step $\ell = m + 1$, stop and reject all hypotheses.

Equivalently, the above procedure can be defined as rejecting $H_{0,i}$ whenever

$$\max_{p_{(\ell)} \leq p_i} \{q_\ell(p_{(\ell)})/\ell\} \leq \gamma.$$

The rationale behind this algorithm is that at step ℓ , when rejecting the ℓ null hypotheses corresponding to the p -values smaller than $p_{(\ell)}$, the number of false discoveries behaves as if it was stochastically dominated by a binomial variable of parameter $(m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1, p_{(\ell)})$. Hence, by controlling the $(1 - \alpha)$ -quantile of the latter binomial variable at level γ^ℓ , the $(1 - \alpha)$ -quantile of the FDP should be controlled by γ . The rigorous proof of the corresponding FDP control is given in Section 5, see Corollary 5.4. Finally, when controlling the median of the FDP, this procedure is related to the recent adaptive procedure of [26], as discussed in Section 6.3.

2. Key concepts and tools

2.1. Model assumptions

Throughout this paper, we will consider several models. Each model corresponds to a specific assumption on the p -value family $\mathbf{p} = \{p_i, 1 \leq i \leq m\}$ distribution. The first model, called the “independent model” is defined as follows:

$$\mathcal{P}^I = \left\{ P \in \mathcal{P} : (p_i(X))_{i \in \mathcal{H}_0(P)} \text{ is a family of mutually independent variables and } (p_i(X))_{i \in \mathcal{H}_0(P)} \text{ is independent of } (p_i(X))_{i \in \mathcal{H}_1(P)} \right\}. \quad (9)$$

The second model uses a particular notion of positive dependence between the p -values, called “weak positive regression dependency” (in short, “weak PRDS”), which is a slightly weaker version of the PRDS assumption of [8]. To introduce the weak PRDS property, let us define a subset $D \subset [0, 1]^m$ as *nondecreasing* if for all $q, q' \in [0, 1]^m$ such that $\forall i \in \{1, \dots, m\}, q_i \leq q'_i$, we have $q' \in D$ when $q \in D$.

Definition 2.1 (Weak PRDS p -value family). *The family \mathbf{p} is said to be weak PRDS on $\mathcal{H}_0(P)$ if for any $i_0 \in \mathcal{H}_0(P)$ and for any measurable nondecreasing set $D \subset [0, 1]^m$, the function $u \mapsto \mathbb{P}(\mathbf{p} \in D \mid p_{i_0} \leq u)$ is nondecreasing on the set $\{u \in [0, 1] : \mathbb{P}(p_{i_0} \leq u) > 0\}$.*

The only difference between the weak PRDS assumption and the “regular” PRDS assumption defined in [8] is that the latter assumes “ $u \mapsto \mathbb{P}(\mathbf{p} \in D \mid p_{i_0} = u)$ nondecreasing”, instead of “ $u \mapsto \mathbb{P}(\mathbf{p} \in D \mid p_{i_0} \leq u)$ nondecreasing”. Weak PRDS is a weaker assumption, as shown for instance in the proof of Proposition 3.6 in [12]. We can now define the second model, where the p -values have weak PRDS dependency:

$$\mathcal{P}^{pos} = \{P \in \mathcal{P} : \mathbf{p}(X) \text{ is weak PRDS on } \mathcal{H}_0(P)\}. \quad (10)$$

It is not difficult to see that $\mathcal{P}^I \subset \mathcal{P}^{pos}$ because when $P \in \mathcal{P}^I$, p_{i_0} is independent of $(p_i)_{i \neq i_0}$ for any $i_0 \in \mathcal{H}_0(P)$. Furthermore, we refer to the general case of $P \in \mathcal{P}$ (without any additional restriction) as the “arbitrary dependence case”.

As an illustration, in the one-sided Gaussian testing framework of Example 1.2, the PRDS assumption (regular and thus also weak) is satisfied as soon as the covariance matrix $\Sigma(P)$ has nonnegative entries, as shown in [8] (note that this is not true anymore for two-sided tests, as proved in the latter reference).

2.2. Dirac configurations

If we want to check whether a procedure satisfies a type I error rate control (7), particularly simple p -value distributions (or “configurations”) are as follows:

- “Dirac configurations”: the p -values of $\mathcal{H}_1(P)$ are equal to zero (without any assumption on the p -values of $\mathcal{H}_0(P)$);
- “Dirac-uniform configuration” (see [24]): the Dirac configuration for which the variables $(p_i)_{i \in \mathcal{H}_0(P)}$ are i.i.d. uniform.

These configurations can be seen as the asymptotic p -value family distribution where the sample size available to perform each test tends to infinity, while the number m of tests is kept fixed (see the examples of Section 1.2). This situation does not fall into the classical multiple testing framework where the number of tests is much larger than the sample size. Besides, there is no multiple testing problem in these configurations because the true nulls are perfectly separated from the false null (almost surely). However, these special configurations are still interesting, because they sometimes have the property to be the distributions for which the type I error rate is the largest. In that case, they are called the “least favorable configurations” (see [24]). This generally requires that the multiple testing procedure and the error rate under consideration have special monotonic properties (see [23, 48]). In this case, proving the type I error rate control for the Dirac configurations is sufficient to state (7) and thus appears to be very useful.

2.3. Algorithms

To derive (7), a generic method that emerged from the multiple testing literature is as follows:

1. start with a family $(R_\kappa)_\kappa$ of procedures depending on an external parameter κ ;

2. find a set of values of κ for which R_κ satisfies (7);
3. take among these values the κ that makes R_κ the “largest”.

The latter is designed to maintain the control of the type I error rate while maximizing the rejection set. As we will see in Section 3 (κ is a threshold t), Section 4 (κ is a subset \mathcal{C} of \mathcal{H}) and Section 5 (κ is a rejection number ℓ), this gives rise to the so-called “step-up” and “step-down” algorithms, which are very classical instances of type I error rate controlling procedures.

2.4. Adaptive control

A way to increase the power of type I error rate controlling procedures is to learn (from the data) part of the unknown distribution P in order to make more rejections. This approach is called “adaptive type I error rate control”. Since the resulting procedure uses the data twice, the main challenge is often to show that it maintains the type I error control (7). In this paper, we will discuss adaptivity with respect to the parameter $m_0(P) = |\mathcal{H}_0(P)|$ for the FDR in Section 3.3. The procedures presented in Section 4 (resp. Section 5) for controlling the k -FWER (resp. FDP) will be also adaptive to $m_0(P)$, but in a maybe more implicit way. Some of them will be additionally adaptive with respect to the dependency structure between the p -values. Let us finally note that some other work studied the adaptivity to the alternative distributions of the p -values (see [62, 49, 47]).

3. FDR control

After the seminal work of Benjamini and Hochberg [5], many studies have investigated the FDR controlling issue. We provide in this section a survey of some of these approaches.

3.1. Thresholding based procedures

Let us start from thresholding type multiple-testing procedures

$$R_t = \{1 \leq i \leq m : p_i \leq t(\mathbf{p})\},$$

with a threshold $t(\cdot) \in [0, 1]$ possibly depending on the p -values. We want to find t such that the corresponding multiple testing procedure R_t controls the FDR at level α under the model \mathcal{P}^{pos} , by following the general method explained in Section 2.3. We start with the following simple decomposition of the false discovery rate of R_t :

$$\text{FDR}(R_t, P) = \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t(\mathbf{p})\}}{\alpha \widehat{\mathbb{G}}(\mathbf{p}, t(\mathbf{p})) \vee (\alpha/m)} \right], \quad (11)$$

where $\widehat{\mathbb{G}}(\mathbf{p}, u) = m^{-1} \sum_{i=1}^m \mathbf{1}\{p_i \leq u\}$ denotes the empirical c.d.f. of the p -value family $\mathbf{p} = \{p_i, 1 \leq i \leq m\}$ taken at a threshold $u \in [0, 1]$.

In order to upper-bound the expectation in the RHS of (11), let us consider the following informal reasoning: if t and $\widehat{\mathbb{G}}$ were deterministic, this expectation would be smaller than $t/(\alpha \widehat{\mathbb{G}}(\mathbf{p}, t))$

and thus smaller than 1 by taking a threshold t such that $t \leq \alpha \widehat{\mathbb{G}}(\mathbf{p}, t)$. This motivates the introduction of the following set of thresholds:

$$\mathcal{T}(\mathbf{p}) = \{u \in [0, 1] : \widehat{\mathbb{G}}(\mathbf{p}, u) \geq u/\alpha\}. \quad (12)$$

With different notation, the latter was introduced in [12, 23]. Here, any threshold $t \in \mathcal{T}(\mathbf{p})$ is said “self-consistent” because it corresponds to a procedure $R_t = \{1 \leq i \leq m : p_i \leq t\}$ which is “self-consistent” according to the definition given in [12], that is, $R_t \subset \{1 \leq i \leq m : p_i \leq \alpha |R_t|/m\}$. It is important to note that the set $\mathcal{T}(\mathbf{p})$ only depends on the p -value family (and on α) so that self-consistent thresholds can be easily chosen in practice. As an illustration, we depict the set $\mathcal{T}(\mathbf{p})$ in Figure 1 for a particular realization of the p -value family.

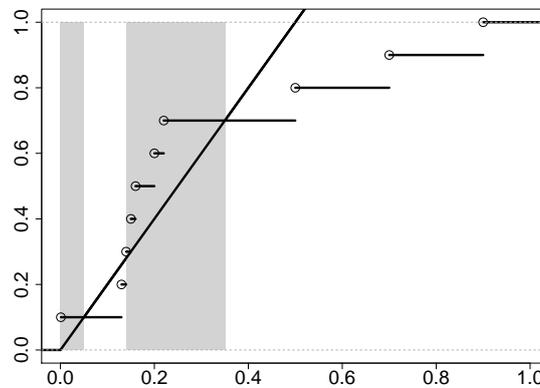


FIGURE 1. The p -value e.c.d.f. $\widehat{\mathbb{G}}(\mathbf{p}, u)$ and u/α are plotted as functions of $u \in [0, 1]$. The points u belonging to the set $\mathcal{T}(\mathbf{p})$ lie on the X-axis of the gray area. $m = 10$; $\alpha = 0.5$.

Now, let us choose a self-consistent threshold $t(\mathbf{p}) \in \mathcal{T}(\mathbf{p})$. By using the decomposition (11), we obtain the following upper-bound:

$$\text{FDR}(R_t, P) \leq \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t(\mathbf{p})\}}{t(\mathbf{p}) \vee (\alpha/m)} \right] \leq \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t(\mathbf{p})\}}{t(\mathbf{p})} \right], \quad (13)$$

with the convention $\frac{0}{0} = 0$. Since by (2), we have $p_i(x) > 0$ for P -almost every x when $i \in \mathcal{H}_0(P)$, the denominator inside the expectation of the RHS of (13) can only be zero when the numerator is also zero and therefore when the ratio is zero. Next, the following purely probabilistic lemma holds (see a proof in Appendix A of [12] for instance):

Lemma 3.1. *Let U be a nonnegative random variable which is stochastically lower bounded by a uniform distribution, i.e., $\mathbb{P}(U \leq u) \leq u$ for any $u \in [0, 1]$. Then the following inequality holds:*

$$\mathbb{E} \left[\frac{\mathbf{1}\{U \leq V\}}{V} \right] \leq 1, \quad (14)$$

for any nonnegative random variable V satisfying either of the two following conditions:

- (i) $V = g(U)$ where $g : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ is non-increasing,
- (ii) the conditional distribution of V conditionally on $U \leq u$ is stochastically decreasing in u , that is, $\forall v \geq 0, u \mapsto \mathbb{P}(V < v | U \leq u)$ is nondecreasing on $\{u \in [0, 1] : \mathbb{P}(U \leq u) > 0\}$.

A consequence of the previous lemma in combination with (13) is that the FDR is controlled at level $\alpha m_0(P)/m$ as soon as $V = t(\mathbf{p})$ satisfies (ii) with $U = p_i$. For the latter to be true, we should make the distributional assumption $P \in \mathcal{P}^{pos}$ and add the assumption that the threshold $t(\cdot)$ is non-increasing with respect to each p -value, that is, for all $q, q' \in [0, 1]^m$, we have $t(q) \leq t(q')$ as soon as for all $1 \leq i \leq m, q'_i \leq q_i$. By using the latter, we easily check that the set

$$D = \{q \in [0, 1]^m : t(q) < v\}$$

is a nondecreasing measurable set of $[0, 1]^m$, for any $v \geq 0$. Thus, the weak PRDS condition defined in Section 2.1 provides (ii) with $U = p_i$ and $V = t(\mathbf{p})$ and thus also (14). Summing up, we obtained the following result, which appeared in [12]:

Theorem 3.2. Consider a thresholding type multiple testing procedure R_t based on a threshold $t(\cdot)$ satisfying the two following conditions:

- $t(\cdot)$ is self-consistent, i.e., such that for all $q \in [0, 1]^m, t(q) \in \mathcal{T}(q)$ (where $\mathcal{T}(\cdot)$ is defined by (12))
- $t(\cdot)$ is coordinate-wise non-increasing, i.e., satisfying that for all $q, q' \in [0, 1]^m$ with $q'_i \leq q_i$ for all $1 \leq i \leq m$, we have $t(q) \leq t(q')$.

Then, for any $P \in \mathcal{P}^{pos}$, $FDR(R_t, P) \leq \alpha m_0(P)/m \leq \alpha$.

Remark 3.3. If we want to state the FDR control of Theorem 3.2 only for $P \in \mathcal{P}^I$ without using the PRDS property, we can use Lemma 3.1 (i) conditionally on $\mathbf{p}_{-i} = (p_j, j \neq i) \in [0, 1]^{m-1}$, by taking $V = t(U, \mathbf{p}_{-i})$ and $U = p_i$, because p_i is independent of \mathbf{p}_{-i} when $P \in \mathcal{P}^I$.

3.2. Linear step-up procedures

From Theorem 3.2, under the weak PRDS assumption on the p -value dependence structure, any algorithm giving as output a self-consistent and non-increasing threshold $t(\cdot)$ leads to a correct FDR control. As explained in Section 1.6 and Section 2.3, for the same FDR control we want to get a procedure with a rejection set as large as possible. Hence, it is natural to choose the following threshold:

$$t^{su}(\mathbf{p}) = \max\{\mathcal{T}(\mathbf{p})\} \tag{15}$$

$$\begin{aligned} &= \max\{u \in \{\alpha k/m, 0 \leq k \leq m\} : \widehat{\mathbb{G}}(\mathbf{p}, u) \geq u/\alpha\} \\ &= \alpha/m \times \max\{0 \leq k \leq m : p_{(k)} \leq \alpha k/m\}, \end{aligned} \tag{16}$$

where $p_{(1)} \leq \dots \leq p_{(m)}$ ($p_{(0)} = 0$) denote the order statistics of the p -value family. This choice was made in [5] and is usually called *linear step-up* or “Benjamini-Hochberg” thresholding. One should notice that the maximum in (15) exists because the set $\mathcal{T}(\mathbf{p})$ contains 0, is upper-bounded by 1 and because the e.c.d.f. is a non-decreasing function (the right-continuity is not needed). It is also easy to check that the maximum $u = t^{su}(\mathbf{p})$ satisfies the equality $\widehat{\mathbb{G}}(\mathbf{p}, u) = u/\alpha$, so that $t^{su}(\mathbf{p})$ can be seen as the largest crossing point between between $u \mapsto \widehat{\mathbb{G}}(\mathbf{p}, u)$ and $u \mapsto u/\alpha$, see

the left-side of Figure 2. The latter equality also implies that $t^{su}(\mathbf{p}) \in \{\alpha k/m, 0 \leq k \leq m\}$, which, combined with the so-called switching relation

$$m \widehat{\mathbb{G}}(\mathbf{p}, \alpha k/m) \geq k \iff p_{(k)} \leq \alpha k/m,$$

gives rise to the second formulation (16). The latter is illustrated in the right-side of Figure 2. The formulation (16) corresponds to the original expression of [5] while (15) is to be found for instance in [27]. Moreover, it is worth noticing that the procedure $R_{t^{su}}$ using the thresholding $t^{su}(\mathbf{p})$ is also equal to $\{1 \leq i \leq m : p_i \leq t^{su}(\mathbf{p}) \vee \alpha/m\}$, so that it can be interpreted as an intermediate thresholding between the non-corrected procedure using $t = \alpha$ and the Bonferroni procedure using $t = \alpha/m$.

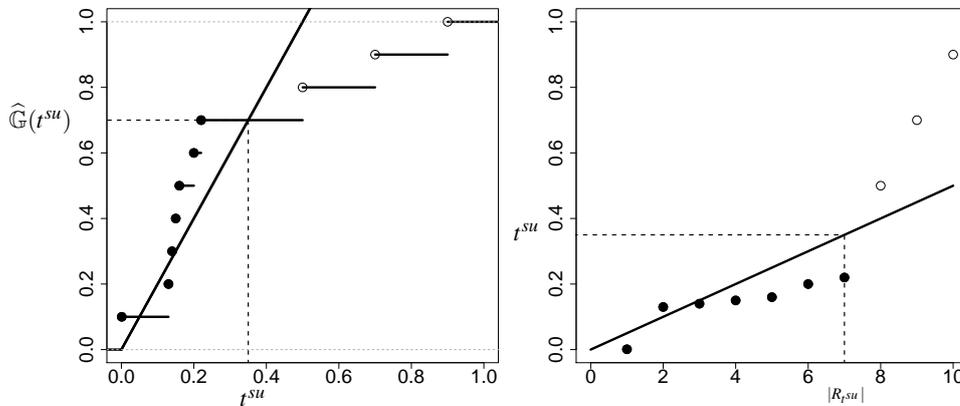


FIGURE 2. The two dual pictorial representations of the Benjamini-Hochberg linear step-up procedure. Left: c.d.f. of the p -values, the solid line has for slope α^{-1} . Right: ordered p -values, the solid line has for slope α/m . In both pictures, the filled points represent p -values that corresponds to the rejected hypotheses. $m = 10$; $\alpha = 0.5$.

Clearly, $t^{su}(\cdot)$ is coordinate-wise non-increasing and self-consistent. Therefore, Theorem 3.2 shows that for any $P \in \mathcal{P}^{pos}$, $\text{FDR}(R_{t^{su}}, P) \leq \alpha m_0(P)/m$. As a matter of fact, as soon as (2) holds with an equality, we can prove that for any $P \in \mathcal{P}^I$, the equality $\text{FDR}(R_{t^{su}}, P) = \alpha m_0(P)/m$ holds, by using a surprisingly direct argument. Let $\mathbf{p}_{0,-i}$ denote the p -value family where p_i has been replaced by 0, and observe that the following statements are equivalent, for any realization of the p -values:

- (i) $p_i \leq t^{su}(\mathbf{p}_{0,-i})$
- (ii) $\widehat{\mathbb{G}}(\mathbf{p}_{0,-i}, t^{su}(\mathbf{p}_{0,-i})) \leq \widehat{\mathbb{G}}(\mathbf{p}, t^{su}(\mathbf{p}_{0,-i}))$
- (iii) $t^{su}(\mathbf{p}_{0,-i})/\alpha \leq \widehat{\mathbb{G}}(\mathbf{p}, t^{su}(\mathbf{p}_{0,-i}))$
- (iv) $t^{su}(\mathbf{p}_{0,-i}) \leq t^{su}(\mathbf{p})$.

The equivalence between (i) and (ii) is straightforward from the definition of $\widehat{\mathbb{G}}(\cdot, \cdot)$. The equivalence between (ii) and (iii) follows from $\widehat{\mathbb{G}}(\mathbf{p}_{0,-i}, t^{su}(\mathbf{p}_{0,-i})) = t^{su}(\mathbf{p}_{0,-i})/\alpha$, because $t = t^{su}(\mathbf{p}_{0,-i})$ is a crossing point between $\widehat{\mathbb{G}}(\mathbf{p}_{0,-i}, t)$ and t/α . The equivalence between (iii) and (iv) comes from the definition of $t^{su}(\mathbf{p})$ together with $t^{su}(\mathbf{p}_{0,-i}) \leq t^{su}(\mathbf{p}) \iff t^{su}(\mathbf{p}_{0,-i}) = t^{su}(\mathbf{p})$, the latter

coming from the non-increasing property of $t^{su}(\cdot)$. As a consequence,

$$\{p_i \leq t^{su}(\mathbf{p}_{0,-i})\} = \{p_i \leq t^{su}(\mathbf{p})\}, \quad (17)$$

with $t^{su}(\mathbf{p}_{0,-i}) = t^{su}(\mathbf{p})$ on these events. Therefore, using (17) and the first decomposition (11) of the FDR, we derive the following equalities:

$$\begin{aligned} \text{FDR}(R_{t^{su}}, P) &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t^{su}(\mathbf{p})\}}{\alpha \widehat{\mathbb{G}}(\mathbf{p}, t^{su}(\mathbf{p})) \vee (\alpha/m)} \right] \\ &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t^{su}(\mathbf{p})\}}{t^{su}(\mathbf{p})} \right] \\ &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t^{su}(\mathbf{p}_{0,-i})\}}{t^{su}(\mathbf{p}_{0,-i})} \right] \\ &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[t^{su}(\mathbf{p}_{0,-i})^{-1} \mathbb{E}(\mathbf{1}\{p_i \leq t^{su}(\mathbf{p}_{0,-i})\} | \mathbf{p}_{0,-i}) \right] \\ &= \alpha m_0(P)/m, \end{aligned}$$

where we assumed in the last equality both that $P \in \mathcal{P}^I$ and condition (2) holds with equality. To sum up, we have proved in this section the following result.

Theorem 3.4. *Consider the linear step-up procedure $R_{t^{su}}$ using the threshold defined in (15). Then, for any $P \in \mathcal{P}^{pos}$, $\text{FDR}(R_{t^{su}}, P) \leq \alpha m_0(P)/m$. Moreover, the latter is an equality if $P \in \mathcal{P}^I$ and (2) holds with equality.*

This theorem is due to [5, 8]. The short proof mentioned above has been independently given in [22, 47, 23]. Theorem 3.4 proves that the inequality “ $\forall P \in \mathcal{P}^{pos}$, $\text{FDR}(R_{t^{su}}, P) \leq \alpha$ ” is sharp as soon as (2) holds with equality and there exists $P \in \mathcal{P}^I$ such that $\mathcal{H}_0(P) = \mathcal{H}$, that is, $\bigcap_{i \in \mathcal{H}} \Theta_{0,i} \cap \mathcal{P}^I \neq \emptyset$.

Other instances of self-consistent procedures include linear “step-up-down” procedures as defined in [50]. Theorem 3.2 establishes that the FDR control also holds for these procedures, as proved in [12, 23].

3.3. Adaptive linear step-up procedures

In this section we denote by $\pi_0(P)$ the proportion $m_0(P)/m$ of hypotheses that are true for P . Since we aim at controlling the FDR at level α and not at level $\alpha\pi_0(P)$, Theorem 3.4 shows that there is a potential power loss when using t^{su} when the proportion $\pi_0(P)$ is small. A first idea is to use the linear step-up procedure at level $\alpha^* = \min(\alpha/\pi_0(P), 1)$, that is, corresponding to the threshold

$$t^*(\mathbf{p}) = \max \{u \in [0, 1] : \widehat{\mathbb{G}}(\mathbf{p}, u) \geq u/\alpha^*\} \quad (18)$$

$$= \max \{u \in [0, 1] : \widehat{\mathbb{G}}(\mathbf{p}, u) \geq u\pi_0(P)/\alpha\}. \quad (19)$$

Note that (18) and (19) are equal because when $\alpha \geq \pi_0(P)$, the maximum is 1 in the two formulas. From Theorem 3.4, threshold (19) provides a FDR smaller than $\alpha^*\pi_0(P) \leq \alpha$ for $P \in \mathcal{P}^{pos}$ and

a FDR equal to α when $P \in \mathcal{P}^I$, (2) holds with equality and $\alpha \leq \pi_0(P)$. Unfortunately, since P is unknown, so is $\pi_0(P)$ and thus the threshold (19) is an unobservable ‘‘oracle’’ threshold.

An interesting challenge is to estimate $\pi_0(P)$ within (19) while still rigorously controlling the FDR at level α , despite the additional fluctuations added by the $\pi_0(P)$ -estimation. This problem, called $\pi_0(P)$ -adaptive FDR control, has received a growing attention in the last decade, see e.g. [6, 56, 9, 28, 7, 41, 51, 13]. To investigate this issue, a natural idea is to consider a modified linear step-procedure using the threshold

$$t_f^{su}(\mathbf{p}) = \max \{u \in [0, 1] : \widehat{\mathbb{G}}(\mathbf{p}, u) \geq u/(\alpha f(\mathbf{p}))\}. \quad (20)$$

where $f(\mathbf{p}) > 0$ is an estimator of $(\pi_0(P))^{-1}$ to be chosen. The latter is called *adaptive linear step-up procedure*. It is sometimes additionally said ‘‘plug in’’, because (20) corresponds to (19) in which we have ‘‘plugged’’ an estimator of $(\pi_0(P))^{-1}$. Other types of adaptive procedures can be defined, see Remark 3.6 below.

We describe now a way to choose f so that the control $\text{FDR}(R_{t_f^{su}}, P) \leq \alpha$ still holds. However, we only focus on the case where the p -values are independent, that is, $P \in \mathcal{P}^I$. This restriction is usual in studies providing an adaptive FDR control. First, to keep the non-increasing property of the threshold $t_f^{su}(\cdot)$, we assume that $f(\cdot)$ is coordinate-wise non-increasing. Second, using techniques similar to those of Section 3.2, we can write for any $P \in \mathcal{P}^I$,

$$\begin{aligned} \text{FDR}(R_{t_f^{su}}, P) &\leq \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t_f^{su}(\mathbf{p})\}}{t_f^{su}(\mathbf{p})} f(\mathbf{p}) \right] \\ &\leq \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t_f^{su}(\mathbf{p})\}}{t_f^{su}(\mathbf{p})} f(\mathbf{p}_{0,-i}) \right] \\ &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[f(\mathbf{p}_{0,-i}) \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq t_f^{su}(\mathbf{p})\}}{t_f^{su}(\mathbf{p})} \middle| \mathbf{p}_{0,-i} \right] \right] \\ &\leq \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E}[f(\mathbf{p}_{0,-i})], \end{aligned} \quad (21)$$

where we used Lemma 14 (i) in the last inequality (conditionally on the p -values of $(p_j, j \neq i)$, because f is coordinate-wise non-increasing). Additionally assuming that $f(\cdot)$ is permutation invariant, we can upper-bound the RHS of (21) by using the Dirac-uniform configuration because $f(\cdot)$ is non-increasing. This gives rise to the following result.

Theorem 3.5. *Consider the adaptive linear step-up procedure $R_{t_f^{su}}$ with a threshold defined in (20) using a $(\pi_0(P))^{-1}$ -estimator f satisfying the following properties:*

- $f(\cdot)$ is coordinate-wise non-increasing, that is, for all $q, q' \in [0, 1]^m$ with for all $1 \leq i \leq m$, $q'_i \leq q_i$, we have $f(q) \leq f(q')$;
- $f(\cdot)$ is permutation invariant, that is, for any permutation σ of $\{1, \dots, m\}$, $\forall q \in [0, 1]^m$, $f(q_1, \dots, q_m) = f(q_{\sigma(1)}, \dots, q_{\sigma(m)})$;
- f satisfies

$$\forall m_0 \in \{1, \dots, m\}, \mathbb{E}_{\mathbf{p} \sim DU(m_0-1, m)}(f(\mathbf{p})) \leq m/m_0, \quad (22)$$

where $DU(k, m)$ denotes the Dirac-uniform distribution on $[0, 1]^m$ for which the k first coordinates are i.i.d. uniform on $(0, 1)$ and the remaining coordinates are equal to 0.

Then, for any $P \in \mathcal{P}^I$, $FDR(R_{f^m}, P) \leq \alpha$.

The method leading to the upper-bound (21) was investigated in [7] and described latter in detail in [13]. The simpler result presented in Theorem 3.5 appeared in [13]. It uses the Dirac-uniform configuration as a least favorable configuration for the FDR. This kind of reasoning has been also used in [23].

Let us now consider the problem of finding a ‘‘correct’’ estimator f of $(\pi_0(P))^{-1}$. This issue has an interest in its own right and many studies investigated it since the first attempt in [52] (see for instance the references in [14]). Here, we only deal with this problem from the FDR control point of view, by providing two families of estimators that satisfy the assumptions of Theorem 3.5. First, define the ‘‘Storey-type’’ estimators, which are of the form

$$f_1(\mathbf{p}) = \frac{m(1 - \lambda)}{\sum_{i=1}^m \mathbf{1}\{p_i > \lambda\} + 1},$$

for $\lambda \in (0, 1)$ (λ not depending on \mathbf{p}). It is clearly non-increasing and permutation invariant. Moreover, we can check that f_1 satisfies (22): for any $m_0 \in \{1, \dots, m\}$, considering $(U_i)_{1 \leq i \leq m_0-1}$ i.i.d. uniform on $(0, 1)$,

$$\mathbb{E}_{\mathbf{p} \sim DU(m_0-1, m)}(f_1(\mathbf{p})) = \frac{m}{m_0} \mathbb{E} \left[\frac{m_0(1 - \lambda)}{\sum_{i=1}^{m_0-1} \mathbf{1}\{U_i > \lambda\} + 1} \right] \leq \frac{m}{m_0},$$

because for any $k \geq 2$, $q \in (0, 1)$ and for Y having a binomial distribution with parameters $(k-1, q)$, we have $\mathbb{E}((1+Y)^{-1}) \leq (qk)^{-1}$, as stated e.g. in [7]. This type of estimator has been introduced in [55] and proved to lead to a correct FDR control in [56, 7].

The second family of estimators satisfying the assumptions of Theorem 3.5 is the ‘‘quantile-type’’ family, defined by

$$f_2(\mathbf{p}) = \frac{m(1 - p_{(k_0)})}{m - k_0 + 1},$$

for $k_0 \in \{1, \dots, m\}$ (k_0 not depending on \mathbf{p}). The latter may be seen as Storey-type estimators using a data-dependent $\lambda = p_{(k_0)}$. Clearly, $f_2(\cdot)$ is non-increasing and permutation-invariant. Additionally, $f_2(\cdot)$ enjoys (22) because for any $m_0 \in \{1, \dots, m\}$, considering $(U_i)_{1 \leq i \leq m_0-1}$ i.i.d. uniform on $(0, 1)$ ordered as $U_{(1)} \leq \dots \leq U_{(m_0-1)}$,

$$\begin{aligned} \mathbb{E}_{\mathbf{p} \sim DU(m_0-1, m)}(f_2(\mathbf{p})) &= \mathbb{E} \left[\frac{m(1 - U_{(k_0-m+m_0-1)})}{m - k_0 + 1} \right] = \frac{m(1 - \mathbb{E}[U_{(k_0-m+m_0-1)}])}{m - k_0 + 1} \\ &= \frac{m(1 - (k_0 - m + m_0 - 1)_+/m_0)}{m - k_0 + 1} \leq \frac{m}{m_0}, \end{aligned}$$

by using the convention $U_{(j)} = 0$ when $j \leq 0$. These quantile type estimators have been proved to lead to a correct FDR control in [7]. The simple proof above was given in [13].

Which choice should we make for λ or k_0 ? Using extensive simulations (including other type of adaptive procedures), it was recommended in [13] to choose as estimator f_1 with λ close to α , because the corresponding procedure shows a ‘‘good’’ power under independence while it maintains a correct FDR control under positive dependencies (in the equi-correlated Gaussian one-sided model described in Example 1.2). Obviously, a ‘‘dynamic’’ choice of λ (i.e., using

the data) can increase the accuracy of the $(\pi_0(P))^{-1}$ estimation and thus should lead to a better procedure. However, proving that the corresponding FDR control remains valid in this case is an open issue to our knowledge. Also, outside the case of the particular equi-correlated Gaussian dependence structure, very little is known about adaptive FDR control.

Remark 3.6. Some authors have proposed adaptive procedures that are not of the “plug-in” form (20). For instance, we can define the class of “one-stage step-up adaptive procedures”, for which the threshold takes the form $t^{os}(\mathbf{p}) = \max \{u \in [0, 1] : \widehat{\mathbb{G}}(\mathbf{p}, u) \geq r_\alpha(u)\}$, where $r_\alpha(\cdot)$ is a non-decreasing function that depends neither on \mathbf{p} nor on $\pi_0(P)$, see, e.g., [41, 23, 13]. As an illustration, Blanchard and Roquain (2009) have introduced the curve defined by $r_\alpha(t) = (1 + m^{-1})t/(t + \alpha(1 - \alpha))$ if $t \leq \alpha$ and $r_\alpha(t) = +\infty$ otherwise, see [13]. They have proved that the corresponding step-up procedure $R_{t^{os}}$ controls the FDR at level α in the independent model (by using the property of Lemma 14 (i)). Furthermore, Finner et al. (2009) have introduced the “asymptotically optimal rejection curve” (AORC) defined by $r_\alpha(t) = t/(\alpha + t(1 - \alpha))$, see [23]. By contrast with the framework of the present paper, they considered the FDR control only in an asymptotic manner where the number m of hypotheses tends to infinity. They have proved that the AORC enjoys the following (asymptotic) optimality property: while several adaptive procedures based on the AORC provide a valid asymptotic FDR control (under independence), the AORC maximizes the asymptotic power among broad classes of adaptive procedures that asymptotically control the FDR, see Theorem 5.1, 5.3 and 5.5 in [23].

3.4. Case of arbitrary dependencies

Many corrections of the linear step-up procedure are available to maintain the FDR control when the p -value family has arbitrary and unknown dependencies. We describe here the so-called “Occam’s hammer” approach presented in [11]. Surprisingly, it allows to recover and extend the well-known “Benjamini-Yekutieli” correction [8] by only using Fubini’s theorem. Let us consider

$$t^{\beta su}(\mathbf{p}) = \max \{u \in [0, 1] : \widehat{\mathbb{G}}(\mathbf{p}, \beta(u)) \geq u/\alpha\} \quad (23)$$

$$\begin{aligned} &= \max \{u \in \{\alpha k/m, 1 \leq k \leq m\} : \widehat{\mathbb{G}}(\mathbf{p}, \beta(u)) \geq u/\alpha\} \\ &= \alpha/m \times \max \{0 \leq k \leq m : p_{(k)} \leq \beta(\alpha k/m)\}, \end{aligned} \quad (24)$$

for a non-decreasing function $\beta : \mathbb{R}^+ \rightarrow \mathbb{R}^+$. Then the FDR of $R_{\beta(t^{\beta su})}$ can be written as follows: for any $P \in \mathcal{P}$,

$$\begin{aligned} \text{FDR}(R_{\beta(t^{\beta su})}, P) &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\frac{\mathbf{1}\{p_i \leq \beta(t^{\beta su}(\mathbf{p}))\}}{t^{\beta su}(\mathbf{p})} \right] \\ &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{E} \left[\mathbf{1}\{p_i \leq \beta(t^{\beta su}(\mathbf{p}))\} \int_0^{+\infty} u^{-2} \mathbf{1}\{t^{\beta su}(\mathbf{p}) \leq u\} du \right]. \end{aligned}$$

Next, using Fubini's theorem, we obtain

$$\begin{aligned}
\text{FDR}(R_{\beta(t^{\beta su})}, P) &= \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \int_0^{+\infty} u^{-2} \mathbb{E}[\mathbf{1}\{t^{\beta su}(\mathbf{p}) \leq u\} \mathbf{1}\{p_i \leq \beta(t^{\beta su}(\mathbf{p}))\}] du \\
&\leq \alpha m^{-1} \sum_{i \in \mathcal{H}_0(P)} \int_0^{+\infty} u^{-2} \mathbb{P}(p_i \leq \beta(u)) du \\
&= \alpha \frac{m_0(P)}{m} \int_0^{+\infty} u^{-2} \beta(u) du.
\end{aligned} \tag{25}$$

Therefore, choosing any non-decreasing function β such that $\int_0^{+\infty} u^{-2} \beta(u) du = 1$ provides a valid FDR control. This leads to the following result:

Theorem 3.7. Consider a function $\beta : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ of the following form: for all $u \geq 0$,

$$\beta(u) = \sum_{i: 1 \leq i \leq m, \alpha i/m \leq u} (\alpha i/m) v_i, \tag{26}$$

where the v_i s are nonnegative with $v_1 + \dots + v_m = 1$. Consider the step-up procedure $R_{\beta(t^{\beta su})}$ using $t^{\beta su}$ defined by (23). Then for any $P \in \mathcal{P}$, $\text{FDR}(R_{\beta(t^{\beta su})}, P) \leq \alpha m_0(P)/m$.

Note that the function β defined by (26) takes the value $(\alpha/m)v_1 + \dots + (\alpha i/m)v_i$ in each $u = \alpha i/m$ and is constant on each interval $(\alpha i/m, \alpha(i+1)/m)$ and on (α, ∞) . Thus, it always satisfies that $\beta(u) \leq u$, for any $u \geq 0$. This means that the procedure $R_{\beta(t^{\beta su})}$ rejects always less hypotheses than the linear step-up procedure $R_{t^{su}}$. Therefore, while $R_{\beta(t^{\beta su})}$ provides a FDR control under no assumption about the p -value dependency structure, it is substantially more conservative than $R_{t^{su}}$ under weak PRDS dependencies between the p -values.

As an illustration, taking $v_i = i^{-1} \delta^{-1}$ for $\delta = 1 + 1/2 + \dots + 1/m$, we obtain $\beta(\alpha i/m) = \delta^{-1} \alpha i/m$, which corresponds to the linear step-up procedure, except that the level α has been divided by $\delta \simeq \log(m)$. This is the so-called Benjamini-Yekutieli procedure proposed in [8]. Theorem 3.7 thus recovers Theorem 1.3 of [8]. We mention another example, maybe less classical, to illustrate the flexibility of the choice of β in Theorem 3.7. By taking $v_{m/2} = 1$ and $v_i = 0$ for $i \neq m/2$ (assuming that $m/2$ is an integer), we obtain $\beta(\alpha i/m) = (\alpha/2) \mathbf{1}\{i \geq m/2\}$. In that case, the final procedure $R_{\beta(t^{\beta su})}$ rejects the hypotheses corresponding to p -values smaller than $\alpha/2$ if $2p_{(m/2)} \leq \alpha$ and rejects no hypothesis otherwise. Theorem 3.7 ensures that this procedure also controls the FDR, under no assumption on the model dependency. Many other choices of β are given in Section 4.2.1 of [12].

Finally, let us underline that any FDR control valid under arbitrary dependency suffers from a lack of interpretability for the underlying FDP, as discussed in Section 6.2.

Remark 3.8 (Sharpness of the bound in Theorem 3.7). In Lemma 3.1 (ii) of [36] (see also [31]), a specifically crafted p -value distribution was built on $[0, 1]^m$ (depending on β) for which the FDR of $R_{\beta(t^{\beta su})}$ is equal to α (and $m_0(P) = m$). If the underlying model \mathcal{P} is such that $(p_i(X))_{1 \leq i \leq m}$ can have this very specific distribution for some $P \in \mathcal{P}$, the inequality " $P \in \mathcal{P}$, $\text{FDR}(R_{\beta(t^{\beta su})}, P) \leq \alpha$ " in Theorem 3.7 is sharp. However, for a "realistic" model \mathcal{P} , this p -value distribution is rarely attained because it assumes quite unrealistic dependencies between the p -values. Related to that, several simulation experiments showed that the standard LSU procedure still provides a good FDR

control under “realistic” dependencies, see e.g. [21, 35]. This means that the corrections defined in this section are generally very conservative for real-life data, because their actually achieved FDR is much smaller than $\alpha m_0(P)/m$. Finally, another drawback of the bound of Theorem 3.7 is that it is much smaller than α when $\pi_0(P) = m_0(P)/m$ is small. To investigate this problem, we can think to apply techniques similar to those of Section 3.3. However, the problem of adaptive FDR control is much more challenging under arbitrary dependency. The few results that are available in this framework are very conservative, see [13].

Remark 3.9 (Aggregation of dependent p -values). Consider Theorem 3.7 in the particular case where all p -values test the same null hypothesis, that is $\Theta_{0,i} = \Theta_0$ for any i . According to Remark 1.6, we obtain a new test of level α , by rejecting H_0 : “ $P \in \Theta_0$ ” if the procedure $R_{\beta(t\beta_{su})}$ defined in Theorem 3.7 rejects at least one null hypothesis, that is, if there exists $k \geq 1$ such that $p_{(k)} \leq \beta(\alpha k/m)$. As an illustration, taking $v_{\gamma m} = 1$ and $v_i = 0$ for $i \neq \gamma m$, for a given $\gamma \in [0, 1]$ such that $\gamma m \in \{1, \dots, m\}$, we obtain $\beta(\alpha i/m) = (\alpha \gamma) \mathbf{1}\{i \geq \gamma m\}$, which gives rise to a test rejecting H_0 whenever $p_{(\gamma m)} \gamma^{-1} \leq \alpha$. This defines a new global p -value

$$\tilde{p} = \min(p_{(\gamma m)} \gamma^{-1}, 1)$$

for testing H_0 that can be seen as an aggregate of the original p -values. Thus, Theorem 3.7 shows that $\mathbb{P}(\tilde{p} \leq \alpha) \leq \alpha$ under the null, for arbitrary dependencies between the original p -values. Interestingly, this aggregation procedure was independently discovered in [40] in a context where one aims at combining p -values that were obtained by different splits of the original sample. Also note that $\gamma = 1/m$ corresponds to the Bonferroni aggregation procedure. Let us finally discuss the choice $\gamma = 1/2$ (assuming that $m/2$ is an integer). In that case, the aggregated p -value is $\tilde{p} = \min(2 p_{(m/2)}, 1)$. According to Remark 3.8, the factor “2” in the latter is needed in theory but may be over-estimated for a “realistic” distribution of the p -value family. As a matter of fact, van de Wiel et al. (2009) have (theoretically) proved that this factor can be dropped as soon as the p -value family has some underlying multivariate Gaussian dependency structure, see [57].

4. k -FWER control

The methodology presented in this section for controlling the k -FWER under arbitrary dependencies can probably be attributed to many authors, e.g. [33, 63, 44, 45]. Here, we opted for a general presentation which emphasizes the rationale of the mathematical argument. This approach has been sketched in the talk [10] and investigated more deeply in [30] where it is referred to as the “sequential rejection principle”. While the latter point of view allows to obtain elegant proofs, it is also useful for developing new FWER controlling procedures (e.g., hierarchical testing, Schaffer improvement), see [30, 29, 34]. This methodology has been initially developed for the FWER. We propose in Section 4.4 a new extension to the k -FWER.

In this section, for simplicity, we drop the explicit dependence of the multiple testing procedure R w.r.t. \mathbf{p} in the notation. The parameter k is fixed in $\{1, \dots, m\}$.

4.1. Subset-indexed family

As a starting point, we assume that there exists a subset-indexed family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ of multiple testing procedures satisfying the two following assumptions:

- $\mathcal{C} \mapsto R_{\mathcal{C}}$ is non-increasing, that is,

$$\forall \mathcal{C}, \mathcal{C}' \subset \mathcal{H} \text{ such that } \mathcal{C} \subset \mathcal{C}', \text{ we have } R_{\mathcal{C}'} \subset R_{\mathcal{C}}; \quad (\text{NI})$$

- $R_{\mathcal{C}}$ controls the k -FWER when \mathcal{C} is equal to the subset of true null hypotheses, that is,

$$\forall P \in \mathcal{P}, k\text{-FWER}(R_{\mathcal{H}_0(P)}, P) \leq \alpha. \quad (\text{FWC}_0)$$

A natural way of deriving such a family is to take a thresholding-based family of the form

$$R_{\mathcal{C}} = \{1 \leq i \leq m : p_i \leq t_{\mathcal{C}}\}, \quad (27)$$

where $t_{\mathcal{C}} \in [0, 1]$ is a threshold which possibly depends on the data $\mathbf{p} = (p_i)_{1 \leq i \leq m}$. Assumption (NI) then holds as soon as we take $t_{\mathcal{C}}$ non-increasing in \mathcal{C} (if $\mathcal{C} \subset \mathcal{C}'$ then $t_{\mathcal{C}'} \leq t_{\mathcal{C}}$). However, $t_{\mathcal{C}}$ should be carefully chosen in order to ensure (FWC₀), as we discuss below.

A first instance of a thresholding-based family satisfying (NI)-(FWC₀) is the ‘‘Bonferroni family’’ that chooses $t_{\mathcal{C}} = \min(\alpha k / |\mathcal{C}|, 1)$. Condition (FWC₀) results from Markov’s inequality:

$$\mathbb{P}(|\mathcal{H}_0(P) \cap R_{\mathcal{H}_0(P)}| \geq k) \leq k^{-1} \sum_{i \in \mathcal{H}_0(P)} \mathbb{P}(p_i \leq t_{\mathcal{H}_0(P)}) \leq |\mathcal{H}_0(P)| t_{\mathcal{H}_0(P)} / k \leq \alpha.$$

This family is not adaptive w.r.t. the dependence structure of the p -values. As an illustration, when the true p -values are all equal, say, to p_{i_0} , $i_0 \in \mathcal{H}_0(P)$, we have

$$\mathbb{P}(|\mathcal{H}_0(P) \cap R_{\mathcal{H}_0(P)}| \geq k) = \mathbb{P}(|\mathcal{H}_0(P)| \mathbf{1}\{p_{i_0} \leq t_{\mathcal{H}_0(P)}\} \geq k) \leq t_{\mathcal{H}_0(P)}.$$

Thus, under this extreme dependency structure, the Bonferroni threshold $\min(\alpha k / |\mathcal{C}|, 1)$ can be replaced by α (the only case which matters is $|\mathcal{C}| \geq k$, see Remark 4.2 below). Hence, there is a potential loss when using the Bonferroni family. In practice, the Bonferroni family is often used as a ‘‘benchmark family’’ for evaluating the performance of other families.

In order to improve on the Bonferroni family, one can try to choose a threshold $t_{\mathcal{C}}$ that captures the dependencies between the p -values while still satisfying (NI)-(FWC₀). For this, first note that for $R_{\mathcal{C}}$ defined by (27),

$$\begin{aligned} k\text{-FWER}(R_{\mathcal{C}}, P) &= \mathbb{P}(\exists i_1, \dots, i_k \in \mathcal{H}_0(P) : \forall i \in \{i_1, \dots, i_k\}, p_i \leq t_{\mathcal{C}}) \\ &= \mathbb{P}(k\text{-min}\{p_i, i \in \mathcal{H}_0(P)\} \leq t_{\mathcal{C}}), \end{aligned}$$

where $k\text{-min}\{p_i, i \in \mathcal{H}_0(P)\}$ denotes the k -th smallest element of $\{p_i, i \in \mathcal{H}_0(P)\}$. Therefore, a natural choice for $t_{\mathcal{C}}$ is the α -quantile of the distribution of $k\text{-min}\{p_i, i \in \mathcal{C}\}$. However, the latter is generally unknown because the underlying distribution P is unknown. An idea is to approximate it by using a randomized thresholding procedure. This method can be applied when the null hypothesis is invariant under the action of a finite group of transformations of the original observation set \mathcal{X} onto itself (such a transformation can be for instance a permutation or a sign-flipping, see [44, 45, 1, 2]). For a recent and general description of this method, we refer the reader to Theorem 2 of [30] (while [30] have developed this method only for $k = 1$, it can be directly generalized to the case of $k \geq 1$). The resulting family satisfies (NI)-(FWC₀) while it is ‘‘adaptive’’ with respect to the p -value dependence structure, in the sense that $t_{\mathcal{C}} = t_{\mathcal{C}}(\mathbf{p})$ implicitly takes into account the potential relations existing between the p -values.

Remark 4.1. The monotonicity condition introduced in [30] can be rewritten with our notation as follows:

$$\forall \mathcal{C}, \mathcal{C}' \subset \mathcal{H} \text{ such that } \mathcal{C} \subset \mathcal{C}', \text{ we have } R_{\mathcal{C}'} \cap \mathcal{C} \subset R_{\mathcal{C}}. \quad (\text{wNI})$$

Condition (wNI) is weaker than condition (NI). Thus, at first sight, the setting of [30] is more general than ours. The next reasoning shows that the two settings are in fact equivalent. Since the condition (FWC₀) only depends on the set of $R_{\mathcal{C}} \cap \mathcal{C}$ (for $\mathcal{C} = \mathcal{H}_0$), we can add the elements of \mathcal{C}^c in the rejection set $R_{\mathcal{C}}$ while still maintaining (FWC₀) true. Therefore, starting from a subset-indexed family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ satisfying the weaker assumptions (wNI)-(FWC₀), we may define a new subset-indexed family $\{R'_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ satisfying our assumptions (NI)-(FWC₀), by letting $R'_{\mathcal{C}} = R_{\mathcal{C}} \cup \mathcal{C}^c$, and then apply to this family the methodology described in the next sections. Moreover, by anticipating the definition of the FWER-controlling algorithm that will be presented in Section 4.4, we can easily check that the output of this algorithm applied to the family $\{R'_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ is the same than the algorithm of [30] applied to the family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$. As a consequence, our framework covers the original setting of [30].

Remark 4.2. Any subset-indexed family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ satisfying (NI)-(FWC₀) can be modified in the following way: take $\tilde{R}_{\mathcal{C}} = \mathcal{H}$ (reject all hypotheses) when $|\mathcal{C}| < k$ and $\tilde{R}_{\mathcal{C}} = R_{\mathcal{C}}$ otherwise. This maintains the conditions (NI)-(FWC₀), because the k -FWER is always zero when $|\mathcal{H}_0(P)| < k$.

In what follows, we investigate the problem of the k -FWER control once we have fixed a subset-indexed family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ satisfying (NI)-(FWC₀).

4.2. Single-step method

From assumption (FWC₀), the procedure $R_{\mathcal{H}_0(P)}$ using $\mathcal{C} = \mathcal{H}_0(P)$ controls the k -FWER. Clearly, this procedure cannot be used because $\mathcal{H}_0(P)$ depends on the unknown underlying distribution P of the data. We can use instead $R_{\mathcal{C}}$ with $\mathcal{C} = \mathcal{H}$ because, from the two assumptions (NI)-(FWC₀) above, we have $k\text{-FWER}(R_{\mathcal{H}}, P) \leq k\text{-FWER}(R_{\mathcal{H}_0(P)}, P) \leq \alpha$. This implies that $R_{\mathcal{H}}$ always controls the k -FWER at level α . The latter is generally called the *single-step* procedure (associated to the family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$). However, we argue that $R_{\mathcal{H}}$ could be often too conservative w.r.t. $R_{\mathcal{H}_0(P)}$, for the two following reasons:

- $\mathcal{H}_0(P)$ can be much smaller than \mathcal{H} ;
- the way the procedures $\{R_{\mathcal{C}}\}$ have been built implicitly assumed that $\mathcal{C} = \mathcal{H}_0(P)$ and can be very conservative when \mathcal{C} is much larger than \mathcal{H}_0 .

For instance, these behaviors have been extensively discussed in [2] for particular Rademacher-resampled thresholding procedures. Therefore, we seek for a procedure controlling the k -FWER which is “close” to $R_{\mathcal{H}_0(P)}$ and which can be derived from the family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ via a simple algorithm.

4.3. Step-down method for FWER

We present in this section the special case of $k = 1$, following the approach of [44] with the presentation proposed in [10, 30]. Let us denote by $A_{\mathcal{C}}$ the sets $(R_{\mathcal{C}})^c$ of non-rejected hypotheses

for the subset-indexed family. Consider the event

$$\Omega_0 = \{R_{\mathcal{H}_0(P)} \cap \mathcal{H}_0(P) = \emptyset\} = \{\mathcal{H}_0(P) \subset A_{\mathcal{H}_0(P)}\}.$$

By assumption (FWC₀), we have $\mathbb{P}(\Omega_0) \geq 1 - \alpha$. Since from (NI), $A_{\mathcal{C}}$ is non-decreasing in \mathcal{C} , the following holds on Ω_0 : for any $\mathcal{C} \subset \mathcal{H}$,

$$\mathcal{H}_0(P) \subset \mathcal{C} \implies A_{\mathcal{H}_0(P)} \subset A_{\mathcal{C}} \implies \mathcal{H}_0(P) \subset A_{\mathcal{C}}. \quad (28)$$

Thus, on the event Ω_0 , taking $\mathcal{C} = \mathcal{C}_0 = \mathcal{H}$ in (28) gives that $\mathcal{H}_0(P) \subset A_{\mathcal{C}_0}$, which in turn implies $\mathcal{H}_0(P) \subset A_{\mathcal{C}_1}$ by taking $\mathcal{C} = \mathcal{C}_1 = A_{\mathcal{C}_0}$ in (28), and so on. By recursion, this proves the following result:

Theorem 4.3. *Assume that a family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ of multiple testing procedures satisfies conditions (NI) and (FWC₀) and consider the corresponding family of non-rejected hypotheses $\{A_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$. Define $\hat{\mathcal{C}}$ by the following “step-down” recursion:*

- Initialization: $\mathcal{C}_0 = \mathcal{H}$;
- Step $j \geq 1$: let $\mathcal{C}_j = A_{\mathcal{C}_{j-1}}$. If $\mathcal{C}_j = \mathcal{C}_{j-1}$, let $\hat{\mathcal{C}} = \mathcal{C}_j$ and stop. Otherwise go to step $j + 1$;

Then the procedure $R = (\hat{\mathcal{C}})^c$, which also equals $R_{\hat{\mathcal{C}}}$, controls the FWER at level α for any $P \in \mathcal{P}$.

Note that for all $j \geq 0$, we have $\mathcal{C}_{j+1} \subset \mathcal{C}_j$, because $\mathcal{C}_1 \subset \mathcal{C}_0$ and $A_{\mathcal{C}}$ is non-decreasing in \mathcal{C} . Thus, the set of rejected hypotheses can only increase during the step-down algorithm. In particular, the final procedure $\hat{\mathcal{C}}^c = R_{\hat{\mathcal{C}}}$ is always less conservative than the single-step procedure $R_{\mathcal{H}}$, for the same FWER control. Thus, using a step-down algorithm is always more powerful than the single-step method.

Example 4.4 (Bonferroni step-down procedure for FWER control). Theorem 4.3 can be used with the Bonferroni family $R_{\mathcal{C}} = \{1 \leq i \leq m : p_i \leq \alpha/|\mathcal{C}|\}$. In that case, by reordering the p -values $p_{(1)} \leq \dots \leq p_{(m)}$ (with $p_{(0)} = 0$), the corresponding step-down procedure defined in Theorem 4.3 can be reformulated as rejecting the nulls with $p_i \leq \alpha/(m - \hat{\ell} + 1)$, where $\hat{\ell} = \max\{\ell \in \{0, 1, \dots, m\} : \forall \ell' \leq \ell, p_{(\ell')} \leq \alpha/(m - \ell' + 1)\}$. This is the well known step-down *Holm procedure* which was introduced and proved to control the FWER in [33]. By contrast with step-up procedures, the step-down Holm procedure starts from the most significant p -value and stops the first time that a (ordered) p -value exceeds the critical curve. This is illustrated in Figure 3.

4.4. Step-down method for k -FWER

We would like to generalize Theorem 4.3 to the case of the k -FWER. This time, we should consider the event

$$\Omega_0 = \{|R_{\mathcal{H}_0(P)} \cap \mathcal{H}_0(P)| \leq k - 1\} = \{\exists I_0 \subset \mathcal{H}, |I_0| = k - 1 : \mathcal{H}_0(P) \subset A_{\mathcal{H}_0(P)} \cup I_0\},$$

which satisfies by assumption $\mathbb{P}(\Omega_0) \geq 1 - \alpha$. For any subset $\mathcal{C} \subset \mathcal{H}$, let

$$\phi(\mathcal{C}) = \bigcup_{I \subset \mathcal{H}, |I|=k-1} A_{\mathcal{C} \cup I} = \bigcup_{I \subset \mathcal{C}^c, |I| \leq k-1} A_{\mathcal{C} \cup I}. \quad (29)$$

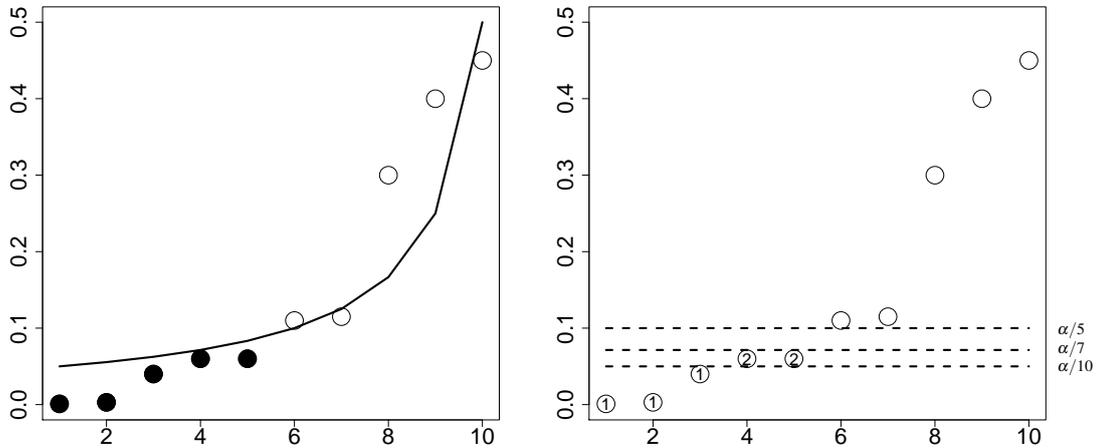


FIGURE 3. Illustration of the two equivalent definitions of Holm's procedure. The left picture is the classical step-down representation: ordered p -values together with the solid curve $\ell \mapsto \alpha/(m - \ell + 1)$. The filled points represent p -values that corresponds to the rejected hypotheses. The right picture illustrates the algorithm of Theorem 4.3: ordered p -values with the three thresholds $\alpha/10$ (step 1), $\alpha/7$ (step 2) and $\alpha/5$ (step 3). For $i \in \{1, 2\}$, the points filled with "i" are rejected in the i th step of the algorithm. Both pictures use the same p -values and $m = 10$; $\alpha = 0.5$.

Then we may prove that the following holds: on the event Ω_0 , for any $\mathcal{C} \subset \mathcal{H}$,

$$\begin{aligned} \exists I \subset \mathcal{H}, |I| = k - 1 : \mathcal{H}_0(P) \subset \mathcal{C} \cup I &\implies \exists I \subset \mathcal{H}, |I| = k - 1 : A_{\mathcal{H}_0(P)} \subset A_{\mathcal{C} \cup I} \subset \phi(\mathcal{C}) \\ &\implies \exists I' \subset \mathcal{H}, |I'| = k - 1 : \mathcal{H}_0(P) \subset \phi(\mathcal{C}) \cup I'. \end{aligned}$$

The first implication holds because $A_{\mathcal{C}}$ is non-decreasing in \mathcal{C} and the second implication holds by considering $I' = I_0$. Thus, on the event Ω_0 , for any $\mathcal{C} \subset \mathcal{H}$,

$$|\mathcal{C}^c \cap \mathcal{H}_0(P)| \leq k - 1 \implies |(\phi(\mathcal{C}))^c \cap \mathcal{H}_0(P)| \leq k - 1.$$

This leads to the following result.

Theorem 4.5. Assume that a family $\{R_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ of multiple testing procedures satisfies conditions (NI) and (FWC₀) and consider the corresponding family of non-rejected hypotheses $\{A_{\mathcal{C}}\}_{\mathcal{C} \subset \mathcal{H}}$ and let ϕ be defined by (29). Define $\hat{\mathcal{C}}$ by the following "step-down" recursion:

- Initialization: $\mathcal{C}_0 = \mathcal{H}$;
 - Step $j \geq 1$: let $\mathcal{C}_j = \phi(\mathcal{C}_{j-1})$. If $\mathcal{C}_j = \mathcal{C}_{j-1}$, let $\hat{\mathcal{C}} = \mathcal{C}_j$ and stop. Otherwise go to step $j + 1$;
- Then the procedure $R = (\hat{\mathcal{C}})^c$, which also equals $(\phi(\hat{\mathcal{C}}))^c = \bigcap_{|I|=k-1} R_{\mathcal{C} \cup I}$, controls the k -FWER at level α for any $P \in \mathcal{P}$.

From (29), $\phi(\cdot)$ is non-decreasing, that is, $\forall \mathcal{C} \subset \mathcal{C}', \phi(\mathcal{C}) \leq \phi(\mathcal{C}')$. As a consequence, we derive from $\mathcal{C}_1 \subset \mathcal{C}_0$ that $\mathcal{C}_j \subset \mathcal{C}_{j-1}$ for all $j \geq 1$. Therefore, the rejection set can only increase at each step of the step-down algorithm. In particular, the final procedure $\hat{\mathcal{C}}^c = \bigcap_{|I|=k-1} R_{\mathcal{C} \cup I}$ is always less conservative than the single step method $R_{\mathcal{H}}$, for the same k -FWER control. Therefore, using the step-down algorithm always leads to a power improvement.

To illustrate Theorem 4.5, let us consider a thresholding-based family of the form $R_{\mathcal{C}} = \{1 \leq i \leq m : p_i \leq t_{\mathcal{C}}\}$ with a non-increasing threshold function $\mathcal{C} \mapsto t_{\mathcal{C}}$ (i.e., such that for $\mathcal{C} \subset \mathcal{C}'$, we have $t_{\mathcal{C}'} \leq t_{\mathcal{C}}$) and such that $\{R_{\mathcal{C}}\}_{\mathcal{C}}$ satisfies (FWC₀). The recursion relation $\mathcal{C}' = \phi(\mathcal{C})$ can be rewritten in that case as follows:

$$\begin{aligned} (\mathcal{C}')^c &= \bigcap_{I \subset \mathcal{C}^c, |I| \leq k-1} R_{\mathcal{C} \cup I} \\ &= \bigcap_{I \subset \mathcal{C}^c, |I| \leq k-1} \{1 \leq i \leq m : p_i \leq t_{\mathcal{C} \cup I}\} \\ &= \{1 \leq i \leq m : p_i \leq \min_{I \subset \mathcal{C}^c, |I| \leq k-1} \{t_{\mathcal{C} \cup I}\}\}. \end{aligned}$$

This recovers the generic step-down method described in Algorithm 2.1 of [45], which was developed in the case where the subset-indexed family is thresholding based.

Example 4.6 (Bonferroni step-down procedure for k -FWER control). When we choose the Bonferroni family, i.e., the threshold family $t_{\mathcal{C}} = \alpha k / |\mathcal{C}|$, we have

$$\min_{I \subset \mathcal{C}^c, |I| \leq k-1} \{t_{\mathcal{C} \cup I}\} = \frac{\alpha k}{m \wedge (|\mathcal{C}| + k - 1)}.$$

Therefore, in terms of the ordered p -values $0 = p_{(0)} \leq p_{(1)} \leq \dots \leq p_{(m)}$, the procedure of Theorem 4.5 can be reformulated as rejecting the null $H_{0,i}$ when $p_i \leq \alpha k / (m \wedge (m - \hat{\ell} + k))$ where $\hat{\ell} = \max\{\ell \in \{0, 1, \dots, m\} : \forall \ell' \leq \ell, p_{(\ell')} \leq \alpha k / (m \wedge (m - \ell' + k))\}$. The latter is the *generalized Holm procedure*, which was introduced and proved to control the k -FWER in [36].

5. FDP control

The problem of controlling the FDP has been investigated in many studies, e.g., [36, 59, 43, 15, 45, 17, 46]. We follow here a methodology proposed by Romano and Wolf (2007), see [45]. They have proposed to use a family $\{S_k\}_k$ of k -FWER controlling procedures and to choose k that ensures that the corresponding rejection number $|S_k|$ is “sufficiently large”. Roughly speaking, choosing k such that $|S_k|$ is larger than $(k-1)/\gamma$ implies that, with high probability,

$$\text{FDP}(S_k, P) = |S_k \cap \mathcal{H}_0(P)| / |S_k| \leq (k-1) / |S_k| \leq \gamma.$$

Obviously, as it is, the above reasoning is not rigorous, because the chosen k depends on the data. Theorem 4.1 (i) of [45] establishes that the latter approach leads to a correct FDP control in the asymptotic setting where the sample size available for each test tends to infinity. This can be seen as a Dirac configuration where each p -value corresponding to false nulls are equal to zero.

In this section, we propose to reformulate this approach by using as index the rejection number instead of k . Roughly speaking, if we choose $\{R_{\ell}\}_{\ell}$ such that each R_{ℓ} controls the $(\gamma\ell + 1)$ -FWER and we choose ℓ such that $|R_{\ell}| \geq \ell$, we obtain that, with high probability,

$$\text{FDP}(R_{\ell}, P) = |R_{\ell} \cap \mathcal{H}_0(P)| / |R_{\ell}| \leq \gamma\ell / |R_{\ell}| \leq \gamma.$$

Similarly to the previous paragraph, this argument is not rigorous because the chosen ℓ depends of the data. The main task of this section is to rationalize this approach. This leads to a general result

(Theorem 5.2 given in Section 5.2), which covers both Theorem 4.1 (i) of [45] in the “Dirac” setting (see Section 5.4) and the earlier result of [36] (see Section 5.3). As additional corollary, we derive the FDP control of the quantile-binomial procedure described in Algorithm 8, when the data are assumed to follow the model \mathcal{P}^I (see Section 5.3).

In this section, the parameter γ is fixed once and for all in $(0, 1)$.

5.1. Family indexed by rejection numbers

Assume that we have at hand a family $\{R_\ell\}_{1 \leq \ell \leq m}$ of multiple testing procedures and a class of distributions $\mathcal{P}' \subset \mathcal{P}$ satisfying the following properties:

- R_ℓ is non-decreasing with respect to ℓ , that is,

$$\forall \ell \in \{1, \dots, m-1\}, R_\ell \subset R_{\ell+1}; \quad (\text{ND})$$

- R_ℓ controls the $(\lfloor \gamma \ell \rfloor + 1)$ -FWER at level α for any $P \in \mathcal{P}'$ such that less than $m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1$ null hypotheses are true, that is,

$$\forall \ell \in \{1, \dots, m\}, \forall P \in \mathcal{P}' \text{ s.t. } |\mathcal{H}_0(P)| \leq m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1, \quad \mathbb{P}(|R_\ell \cap \mathcal{H}_0(P)| \geq \lfloor \gamma \ell \rfloor + 1) \leq \alpha; \quad (\text{FWC})$$

- for any $P \in \mathcal{P}'$, for any $\ell \in \{1, \dots, m\}$, the false rejection number of R_ℓ is independent of the correct rejection numbers of $R_{\ell'}$, for $1 \leq \ell' \leq m$, that is,

$$\forall P \in \mathcal{P}', \forall \ell \in \{1, \dots, m\}, |R_\ell \cap \mathcal{H}_0(P)| \text{ is independent of } \{|R_{\ell'} \cap \mathcal{H}_1(P)|, 1 \leq \ell' \leq m\}. \quad (\text{DA})$$

In condition (FWC), for any $x \geq 0$, $\lfloor x \rfloor$ denotes the largest integer n such that $n \leq x$. Condition (ND) is natural because the index ℓ can be interpreted as a rejection number. It is easy to check in the examples below.

For any $\mathcal{P}' \subset \mathcal{P}$, condition (FWC) is fulfilled by the (single-step or step-down) k -FWER controlling procedures of the previous section when $k = \lfloor \gamma \ell \rfloor + 1$. As a first instance, we can use the (single-step) Bonferroni family R_ℓ using the threshold $\alpha(\lfloor \gamma \ell \rfloor + 1)/m$. Moreover, note that $|\mathcal{H}_0(P)| \leq m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1$ in (FWC), thus we can consider the improved threshold

$$t_\ell^{LR} = \frac{\alpha(\lfloor \gamma \ell \rfloor + 1)}{m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1}. \quad (30)$$

The threshold (30) is slightly larger than the threshold used in Theorem 3.1 of [36] (they used $\lfloor \gamma \ell \rfloor$ instead of $\lfloor \gamma(\ell - 1) \rfloor$ in the denominator). As a second instance, we can substantially improve on the above threshold family when we additionally assume that the distribution P of the data lies in the smaller subset $\mathcal{P}' = \mathcal{P}^I$: for this, note that for any $P \in \mathcal{P}^I$ and for any $t \in [0, 1]$, the variable $|\{i \in \mathcal{H}_0(P) : p_i(X) \leq t\}|$ is stochastically upper-bounded by a binomial distribution of parameters $|\mathcal{H}_0(P)|$ and t , which in turn is stochastically upper-bounded by a binomial distribution of parameters $m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1$ and t . Therefore, choosing the (deterministic) quantile-based threshold family $(t_\ell^Q)_{1 \leq \ell \leq m}$ defined by

$$\begin{aligned} t_\ell^Q &= \max\{t \in [0, 1] : \mathbb{P}(Z > \gamma \ell) \leq \alpha \text{ for } Z \sim \mathcal{B}(m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1, t)\} \\ &= \max\{t \in [0, 1] : q_\ell(t) \leq \gamma \ell\}, \end{aligned} \quad (31)$$

where $q_\ell(\cdot)$ is defined by (8), we obtain a family of thresholding procedures satisfying (FWC) with $\mathcal{P}' = \mathcal{P}^I$. Clearly, since t_ℓ^{LR} in (30) is only based upon Markov's inequality, which is in general not accurate for binomial variables, the threshold family t_ℓ^Q defined by (31) is substantially larger, as illustrated in Figure 4. Interestingly, we can use more elaborate deviation inequalities to obtain thresholds that are better than t_ℓ^{LR} while having a form more explicit than t_ℓ^Q , see Remark 5.1.

Assumption (DA) is a dependence assumption which is typically satisfied in the two following cases:

- each procedure R_ℓ uses a deterministic threshold and the p -values associated to true nulls are independent of the p -values associated to false nulls, for all distributions of \mathcal{P}' , that is,

$$\forall \ell \in \{1, \dots, m\}, R_\ell = \{i \in \{1, \dots, m\} : p_i \leq t_\ell\} \text{ for a deterministic } t_\ell \in [0, 1] ; \quad (\text{DA}') \\ \text{and } \forall P \in \mathcal{P}', (p_i(X))_{i \in \mathcal{H}_0(P)} \text{ is independent of } (p_i(X))_{i \in \mathcal{H}_1(P)}$$

- for all distributions of \mathcal{P}' , the number of correct rejections of each R_ℓ is deterministic, that is,

$$\forall P \in \mathcal{P}', \{ |R_{\ell'} \cap \mathcal{H}_1(P)|, 1 \leq \ell' \leq m \} \text{ is deterministic.} \quad (\text{DA}'')$$

Condition (DA'') is satisfied for instance when $\mathcal{H}_1(P) \subset R_{\ell'}$, for any ℓ' , which is the case for procedures of the form $R_\ell = \{i \in \{1, \dots, m\} : p_i \leq t_\ell(\mathbf{p})\}$ using a possibly data-dependent threshold $t_\ell(\mathbf{p}) \in [0, 1]$, when we assume that the p -values are in the Dirac configuration, that is, when they are equal to zero under the alternative.

Remark 5.1. Using Hoeffding's and Bennett's inequalities (see, e.g., Proposition 2.7 and 2.8 in [39]), we can derive a family of thresholding procedures satisfying (FWC) with $\mathcal{P}' = \mathcal{P}^I$, by using the threshold

$$(t_\ell^Q)'_\ell = \max(t_\ell^{LR}, t_\ell^{Ho}, t_\ell^{Be}), \quad (32)$$

where we let

$$t_\ell^{Ho} = \left(\frac{\lfloor \gamma \ell \rfloor + 1}{m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1} - \left(\frac{\log(1/\alpha)}{2(m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1)} \right)^{1/2} \right) \vee 0 \\ t_\ell^{Be} = \frac{\lfloor \gamma \ell \rfloor + 1}{m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1} h^{-1} \left(\frac{\log(1/\alpha)}{\lfloor \gamma \ell \rfloor + 1} \right),$$

with $h(u) = u - \log(u) - 1$, $u \in (0, 1]$.

5.2. Step-down method

The approach described in this section is an adaptation of the proof of Theorem 3.1 in [36] to our setting. Let us consider a family $\{R_\ell\}_{1 \leq \ell \leq m}$ and a class of distributions $\mathcal{P}' \subset \mathcal{P}$ satisfying (ND)-(FWC)-(DA). We aim at selecting $\ell = \hat{\ell}$ that provides $\forall P \in \mathcal{P}'$, $\text{FDP}(R_{\hat{\ell}}, P) \leq \alpha$.

First note that, by definition of the FDP, we have for any $\ell \in \{1, \dots, m\}$ such that $|R_\ell| = \ell$:

$$\{\text{FDP}(R_\ell, P) > \gamma\} = \{|\mathcal{H}_0(P) \cap R_\ell| > \gamma \ell\} \\ = \{|\mathcal{H}_0(P) \cap R_\ell| \geq \lfloor \gamma \ell \rfloor + 1\} \\ = \{\ell \in \mathcal{L}\}, \quad (33)$$

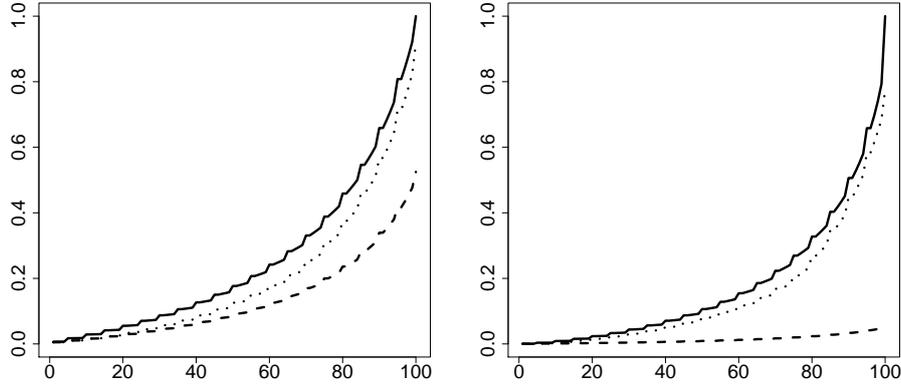


FIGURE 4. Threshold t_ℓ^Q in (31) for model \mathcal{P}^I (solid line), threshold $(t_\ell^Q)'$ in (32) for model \mathcal{P}^I (dotted line) and threshold t_ℓ^{LR} in (30) for model \mathcal{P} (dashed line) in function of $\ell \in \{1, \dots, m\}$. $m = 100$; $\gamma = 0.2$. Left: $\alpha = 0.5$; right: $\alpha = 0.05$.

where $\mathcal{L} = \{\ell \in \{1, \dots, m\} : \ell - |\mathcal{H}_1(P) \cap R_\ell| \geq \lfloor \gamma \ell \rfloor + 1\}$ is a set which only depends on the set $\{|\mathcal{H}_1(P) \cap R_{\ell'}|, 1 \leq \ell' \leq m\}$.

Second, note that for any $\ell \in \{1, \dots, m\}$ such that $|R_\ell| \geq \ell$,

$$\{\ell \in \mathcal{L}\} \subset \{|\mathcal{H}_0(P) \cap R_\ell| \geq \lfloor \gamma \ell \rfloor + 1\}. \quad (34)$$

Let us consider $\ell^* = \min\{\mathcal{L}\}$ (with $\ell^* = m + 1$ when $\mathcal{L} = \emptyset$). From (33) and (34), taking $\hat{\ell} \in \{1, \dots, m\}$ such that $|R_{\hat{\ell}}| = \hat{\ell}$ and such that for any $\ell \leq \hat{\ell}$, $|R_\ell| \geq \ell$, we obtain

$$\begin{aligned} \{\text{FDP}(R_{\hat{\ell}}, P) > \gamma\} &\subset \{\ell^* \leq \hat{\ell}\} \\ &\subset \{|\mathcal{H}_0(P) \cap R_{\ell^*}| \geq \lfloor \gamma \ell^* \rfloor + 1\}. \end{aligned}$$

Moreover, if $\ell^* \geq 2$, by definition of ℓ^* , we have $\ell^* - 1 \notin \mathcal{L}$. Hence, we obtain the following upper-bound for $|\mathcal{H}_0(P)|$:

$$|\mathcal{H}_0(P)| = m - |\mathcal{H}_1(P)| \leq m - |\mathcal{H}_1(P) \cap R_{\ell^*-1}| \leq m - \ell^* + \lfloor \gamma(\ell^* - 1) \rfloor + 1.$$

Since the above bound is also true when $\ell^* = 1$, it holds for any possible value of ℓ^* .

Finally noting that ℓ^* only depends on the variable set $\{|\mathcal{H}_1(P) \cap R_{\ell'}|, 1 \leq \ell' \leq m\}$ and using (FWC)-(DA), we have proved that for any $\ell \in \{1, \dots, m\}$,

$$\begin{aligned} \mathbb{P}(\text{FDP}(R_{\hat{\ell}}, P) > \gamma \mid \ell^* = \ell) &\leq \mathbb{P}(|\mathcal{H}_0(P) \cap R_\ell| \geq \lfloor \gamma \ell \rfloor + 1 \mid \ell^* = \ell) \\ &= \mathbb{P}(|\mathcal{H}_0(P) \cap R_\ell| \geq \lfloor \gamma \ell \rfloor + 1) \\ &\leq \alpha. \end{aligned}$$

Also, the probability $\mathbb{P}(\text{FDP}(R_{\hat{\ell}}, P) > \gamma \mid \ell^* = m + 1)$ is zero, because it is smaller than $\mathbb{P}(\hat{\ell} \in \mathcal{L} \mid \ell^* = m + 1)$. This leads to the following result.

Theorem 5.2. *Assume that there exists a family $\{R_\ell\}_{1 \leq \ell \leq m}$ of multiple testing procedures and a class of distributions $\mathcal{P}' \subset \mathcal{P}$ satisfying the conditions (ND)-(FWC)-(DA) defined in Section 5.1. Consider the procedure $R_{\hat{\ell}}$ where*

$$\hat{\ell} = \max \{ \ell \in \{0, \dots, m\} : \forall \ell' \in \{0, \dots, \ell\}, |R_{\ell'}| \geq \ell' \}, \quad (35)$$

(with the convention $R_0 = \emptyset$). Then $R_{\hat{\ell}}$ controls the FDP in the following sense:

$$\forall P \in \mathcal{P}', \mathbb{P}(\text{FDP}(R_{\hat{\ell}}, P) > \gamma) \leq \alpha. \quad (36)$$

The algorithm performed to find (35) is a step-down algorithm; it starts from small rejection numbers and stops the first time that $|R_\ell|$ is below ℓ . Note that the maximum in (35) is well defined because $\ell = 0$ satisfies $|R_\ell| \geq \ell$. Furthermore, using (ND), relation (35) implies $\hat{\ell} \leq |R_{\hat{\ell}}| \leq |R_{\hat{\ell}+1}| < \hat{\ell} + 1$, so that $|R_{\hat{\ell}}| = |R_{\hat{\ell}+1}| = \hat{\ell}$ holds. As a consequence, the procedure of Theorem 5.2 can be equivalently defined by $R_{\tilde{\ell}}$ where

$$\tilde{\ell} = \min \{ \ell \in \{1, \dots, m+1\} : |R_\ell| \leq \ell - 1 \}, \quad (37)$$

with the convention $R_{m+1} = R_m$ (so that the minimum in (37) is well defined).

5.3. Theorem 3.1 of [36] and the quantile-binomial procedure as corollaries

Going back to the specific setting (DA') described in Section 5.1, we may derive from Theorem 5.2 the following corollary.

Corollary 5.3. *Let us consider the deterministic threshold family $(t_\ell^{LR})_{1 \leq \ell \leq m}$ defined by (30) and consider*

$$\hat{\ell} = \max \{ \ell \in \{0, \dots, m\} : \forall \ell' \in \{0, \dots, \ell\}, p_{(\ell')} \leq t_{\ell'}^{LR} \}, \quad (38)$$

where $0 = p_{(0)} \leq p_{(1)} \leq \dots \leq p_{(m)}$ denote the ordered p -values and by convention $t_0^{LR} = 0$. Then the procedure $R_{\hat{\ell}} = \{i \in \{1, \dots, m\} : p_i \leq t_{\hat{\ell}}^{LR}\}$ satisfies the FDP control (36) for the subset \mathcal{P}' of distributions $P \in \mathcal{P}$ such that the family $(p_i(X))_{i \in \mathcal{H}_0(P)}$ is independent of the family $(p_i(X))_{i \in \mathcal{H}_1(P)}$.

By reproducing the end of the proof of Theorem 5.2 in the particular setting of Corollary 5.3, we may increase a bit the distribution set \mathcal{P}' in Corollary 5.3 to the set of $P \in \mathcal{P}$ such that for any $i \in \mathcal{H}_0(P)$, $\forall u \in [0, 1]$, $\mathbb{P}(p_i(X) \leq u \mid (p_i(X))_{i \in \mathcal{H}_1(P)}) \leq u$. This is the distributional setting of Theorem 3.1 of [36]. Hence, we are able to recover the latter result (with a slight improvement in the threshold family).

Furthermore, if we want to ensure the FDP control (36) only for the smaller distribution set $\mathcal{P}' = \mathcal{P}^I$, we may consider the larger threshold family $(t_\ell^Q)_{1 \leq \ell \leq m}$ defined by (31). This gives rise to the step-down procedure

$$R^Q = \{i \in \{1, \dots, m\} : p_i \leq t_{\hat{\ell}}^Q\}, \quad (39)$$

where $\hat{\ell} = \max \{ \ell \in \{0, \dots, m\} : \forall \ell' \in \{0, \dots, \ell\}, p_{(\ell')} \leq t_{\ell'}^Q \}$ (with $t_0^Q = 0$). The latter is the procedure described in Algorithm 1.7, because $p_{(\ell)} \leq t_{\ell}^Q$ if and only if $q_\ell(p_{(\ell)}) \leq \gamma \ell$, with $q_\ell(\cdot)$ defined by (8). As a consequence, Theorem 5.2 provides the result announced in Section 1.7.

Corollary 5.4. For any $\gamma, \alpha \in (0, 1)$, the quantile-binomial procedure R^Q described in Algorithm 1.7, or equivalently in (39), controls the FDP in the following way:

$$\forall P \in \mathcal{P}^I, \mathbb{P}(FDP(R^Q, P) > \gamma) \leq \alpha.$$

In particular, the median-binomial procedure R^M (using $\alpha = 1/2$) provides that the median of the distribution of $FDP(R^M, P)$ is controlled at level γ for any $P \in \mathcal{P}^I$.

To our knowledge, the above result is a new finding. It establishes a FDP control which is substantially more suitable to the case of independent p -values in comparison with the procedure of [36]. Further comments on this procedure can be found in Section 6.3.

5.4. Theorem 4.1 (i) of [45] as a corollary

In Section 4 of [45], a step-down procedure $S_{\hat{k}}$ is defined from a generic family $\{S_k\}_{1 \leq k \leq m}$ of thresholding based procedures. The latter family is assumed to be such that each S_k controls the k -FWER for $1 \leq k \leq m$ and $S_k \subset S_{k+1}$ for $1 \leq k \leq m-1$. The index \hat{k} is obtained as follows:

$$\hat{k} = \min\{k \in \{1, \dots, m+1\} : \gamma |S_k| < k - \gamma\}, \quad (40)$$

where we use here the convention $S_{m+1} = S_m$ (so that the above set always contains $k = m+1$). Theorem 4.1 (i) of [45] states that $S_{\hat{k}}$ controls the FDP in the asymptotic sense, as the sample size available to perform each test tends to infinity. This can be seen as a (non-asymptotic) FDP control in a Dirac configuration where the p -values corresponding to false nulls are equal to zero. Set under this form, Theorem 4.1 (i) of [45] can be derived from Theorem 5.2.

For this, let $R_\ell = S_{\lfloor \gamma \ell \rfloor + 1}$, for $\ell \in \{1, \dots, m\}$, and note that the family $\{R_\ell\}_{1 \leq \ell \leq m}$ satisfies (ND)-(FWC) and (DA''), by taking the distribution set \mathcal{P}' corresponding to Dirac configurations for the p -values. Hence, Theorem 5.2 establishes the FDP control for the Dirac configurations of the procedure $R_{\tilde{\ell}}$ where $\tilde{\ell}$ is defined by (35), or equivalently by (37). Thus, it only remains to show that the step-down algorithms (40) and (37) lead to the same procedure, that is,

$$R_{\tilde{\ell}} = S_{\hat{k}}.$$

To prove the latter, we establish $\hat{k} = \lfloor \gamma \tilde{\ell} \rfloor + 1$. First, using (37), $\tilde{\ell}$ satisfies $\gamma |S_{\lfloor \gamma \tilde{\ell} \rfloor + 1}| \leq \gamma \tilde{\ell} - \gamma$. Since $\gamma \ell < \lfloor \gamma \ell \rfloor + 1$, we deduce from the definition of \hat{k} that $\lfloor \gamma \tilde{\ell} \rfloor + 1 \geq \hat{k}$. Conversely, by considering the unique integer $\ell \in \{1, \dots, m\}$ satisfying $\hat{k}/\gamma - 1 \leq \ell < \hat{k}/\gamma$ and thus also $\lfloor \gamma \ell \rfloor + 1 = \hat{k}$, we have that for any integer j , $\gamma j < \hat{k} \Rightarrow j \leq \ell$. Applying the latter for $j = \lfloor S_{\hat{k}} \rfloor + 1$, we obtain from $\gamma(\lfloor S_{\hat{k}} \rfloor + 1) < \hat{k}$ that $\lfloor S_{\hat{k}} \rfloor \leq \ell - 1$ and thus $\ell \geq \tilde{\ell}$, by using the definition of $\tilde{\ell}$. This in turn implies $\hat{k} \geq \lfloor \gamma \tilde{\ell} \rfloor + 1$. We thus have proved the following result, which can be seen as Theorem 4.1 (i) of [45] in the Dirac setting.

Corollary 5.5. Assume that there exists a family $\{S_k\}_{1 \leq k \leq m}$ of multiple testing procedures (with the convention $S_{m+1} = S_m$) satisfying

- for each $k \in \{1, \dots, m\}$, S_k is of the form $\{i \in \{1, \dots, m\} : p_i \leq t_k(\mathbf{p})\}$ for a possibly data-dependent threshold $t_k(\cdot) \in [0, 1]$;
- for each $k \in \{1, \dots, m-1\}$, $S_k \subset S_{k+1}$;

- for each $k \in \{1, \dots, m\}$, $\forall P \in \mathcal{P}$, $k\text{-FWER}(S_k, P) \leq \alpha$.

Consider \hat{k} defined in (40) and the subset \mathcal{P}' of distributions $P \in \mathcal{P}$ corresponding to a Dirac configuration, i.e., such that $\forall P \in \mathcal{P}'$, $\forall i \in \mathcal{H}_1(P)$, $p_i(x) = 0$ for P -almost every $x \in \mathcal{X}$. Then we have $\forall P \in \mathcal{P}'$, $\mathbb{P}(\text{FDP}(S_{\hat{k}}, P) > \gamma) \leq \alpha$.

6. Discussion

6.1. Complexity of the k -FWER step-down approach

One major limitation of the k -FWER approach presented in Section 4 is that the computation of $\phi(\cdot)$ in (29) can become cumbersome when k is large because we should consider all subsets I of \mathcal{C}^c of cardinality $k - 1$ (say that $|\mathcal{C}^c| \geq k - 1$). However, we may modify this algorithm by considering only the set I equals to the $k - 1$ indexes of \mathcal{C}^c corresponding to the $k - 1$ largest p -values in $\{p_i, i \in \mathcal{C}^c\}$. As noted in [45], this “streamlined” step-down procedure still controls the k -FWER in the Dirac model where each false null has a p -value equals to zero. The latter is true because in this model, as soon as $|\mathcal{C}^c \cap \mathcal{H}_0(P)| \leq k - 1$, we know that the set $\mathcal{C}^c \cap \mathcal{H}_0(P)$ is included in the set I of indexes corresponding to the $k - 1$ largest p -values in $\{p_i, i \in \mathcal{C}^c\}$ (because the p -values of $\{p_i, i \in \mathcal{C}^c \cap \mathcal{H}_1(P)\}$ are zero). Nevertheless, no proof of this k -FWER control stands without this Dirac assumption.

6.2. FDR control is not FDP control

Since the only interpretable variable is the FDP and not its expectation, controlling the FDR is meaningful only when the FDP concentrates well around the FDR. As the hypothesis number m grows, Neuvial (2008) showed that the latter holds for step-up type procedures when a Donsker type theorem for the e.c.d.f. is valid, so for instance under independence or “weak” dependence, see [41]. However, under some unspecified dependencies, we do not know how the FDP concentrates. For instance, even under a very simple ρ -equi-correlated Gaussian model (corresponding to Example 1.2, where the non-diagonal entries of $\Sigma(P)$ are all equal to ρ), it was shown in [16] that the convergence rate of the FDP to the FDR can be arbitrarily slow when $\rho = \rho_m$ tends to zero as m tends to infinity. Additionally, it was proved in [24] that no concentration phenomenon occurs when ρ is kept fixed with m . Also, as shown in [48], the “sparsity” ($\pi_0(P) = \pi_{0,m}(P)$ tends to 1 as m tends to infinity) is one other feature that can slow down the FDP convergence. Therefore, in all these cases, the FDP convergence is slow and controlling the FDR does not lead to a clear interpretation for the underlying FDP. The latter drawback does not arise while controlling the FDP upper-tail distribution: for instance, the FDP control $\mathbb{P}(\text{FDP} > 0.01) \leq 0.5$ ensures that, with a probability at least 0.5, the FDP is below 0.01, and this interpretation holds whatever the FDP distribution is. However, the FDR stays useful, because this is a simpler criterion for which the controlling methodology is (for now) much more developed in comparison with the FDP controlling methodology.

6.3. Quantile-binomial procedure and relation to previous work

Let us consider the quantile-binomial procedure defined in algorithm 1.7 and the quantile function $q_\ell(\cdot)$ defined by (8). In the particular case where we take $\alpha = 1/2$, the procedure is called the

median-binomial procedure and Corollary 5.4 shows that it controls the median of the FDP at level γ under independence of the p -values. Interestingly, in the ‘‘Gaussian regime’’ where the underlying binomial variable is close to a Gaussian variable (say, γ not too small, many rejections), the median is close to the expectation and thus $q_\ell(t) \simeq (m - \ell + \lfloor \gamma(\ell - 1) \rfloor + 1)t \simeq (m - (1 - \gamma)\ell + 1)t$. Hence, in this case, the median-binomial procedure is close to the step-down procedure using the thresholding $t_\ell = \gamma\ell / (m - (1 - \gamma)\ell + 1)$. As matter of fact, the latter procedure has been recently introduced by Gavrilov et al. (2009) and it has been proved to control the FDR under independence, see [26]. Roughly speaking, the latter may be interpreted in our framework as a ‘‘mean-binomial procedure’’. However, in the Poisson regime (say, γ small, few rejections), the median-binomial procedure can be substantially different from the procedure of Gavrilov et al. (2009). Hence, we should keep in mind that the two procedures do not control the same error rate. These different remarks are illustrated in Figure 5, where we have also reported the Benjamini-Hochberg threshold.

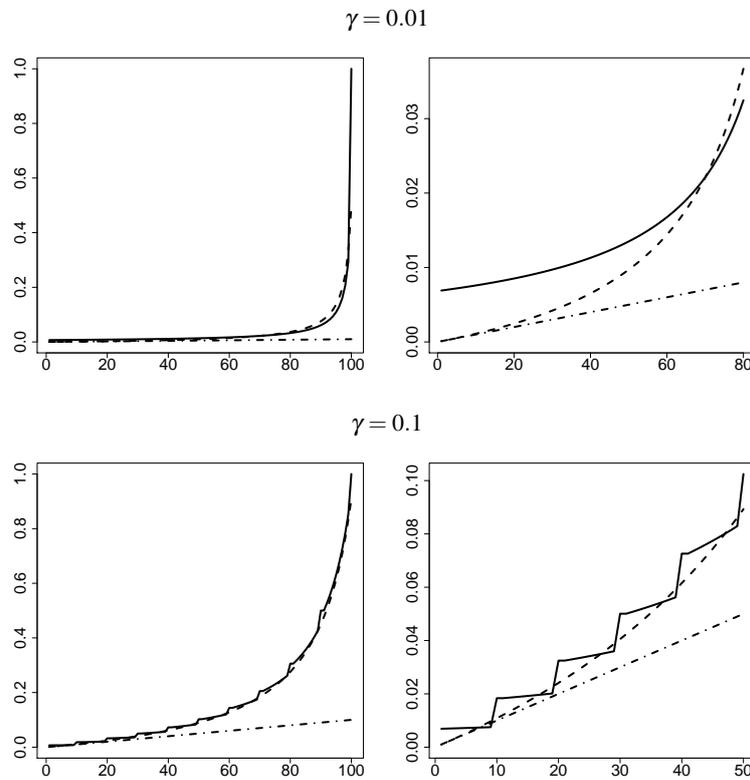


FIGURE 5. Comparison between Benjamini-Hochberg thresholding $t_\ell = \gamma\ell/m$ (dashed-dotted), the Gavrilov et al. thresholding $t_\ell = \gamma\ell / (m - (1 - \gamma)\ell + 1)$ (dashed) and the quantile thresholding t_ℓ^Q defined by (31) with $\alpha = 0.5$ (solid) in function of ℓ . $m = 100$; Top: $\gamma = 0.01$; Bottom: $\gamma = 0.1$. Each right picture is a zoom of the left picture into the region $\ell \in \{1, \dots, 80\}$ (top) or $\ell \in \{1, \dots, 50\}$ (bottom).

6.4. Conclusion

In this paper, we have recovered some of the classical state-of-the-art multiple testing procedures for controlling the FDR, k -FWER and the FDP. Additionally, some new contributions were also given for k -FWER and FDP control, by extending and unifying some previous work of multiple testing literature and by finding a novel procedure, based on the quantiles of the binomial distribution, which controls the FDP under independence.

The type I error rate control research area still has many unsolved issues. Among the major concerns, the FDP control in Section 5 needs a very strong distributional assumption on the test statistics, namely independence or "Dirac" assumption. To our knowledge, no procedure adaptive to dependencies is proved to control the FDP without assuming such a strong requirement. This is a room left for future developments, which would have a strong impact on high-dimensional data analysis.

Acknowledgements

I am very grateful to Lucien Birgé and Gilles Blanchard for their helpful and particularly relevant comments. I also warmly thank Tabea Rebafka and Sylvain Arlot for having carefully read a previous version of this manuscript. This work was supported by the French Agence Nationale de la Recherche (ANR grant references: ANR-09-JCJC-0027-01, ANR-PARCIMONIE, ANR-09-JCJC-0101-01) and the French ministry of foreign and european affairs (EGIDE - PROCOPE project number 21887 NJ).

Appendix A: Defining a p -value from a test statistic

Let us consider the problem of testing a (single) hypothesis $H_0 : "P \in \Theta_0"$ from a test statistic $S(X)$. Assume that H_0 should be rejected for "large" values of $S(X)$. We let $T_P(s) = \mathbb{P}_{X \sim P}(S(X) \geq s)$, $F_P(s) = \mathbb{P}_{X \sim P}(S(X) \leq s)$ and $F_P^{-1}(v) = \min\{s \in \mathbb{R} \cup \{-\infty\} : F_P(s) \geq v\}$. The following result is elementary and can be considered as well known. It is strongly related to Theorem 10.12 in [61], Lemma 3.3.1 in [37] (see also Problem 3.23 therein) and Proposition 1.2 in [17].

Proposition A.1. *The p -value $p(X) = \sup_{P \in \Theta_0} T_P(S(X))$ satisfies the following:*

(i) *$p(X)$ is stochastically lower-bounded by a uniform variable under the null, that is,*

$$\forall P \in \Theta_0, \forall u \in [0, 1], \mathbb{P}_{X \sim P}(p(X) \leq u) \leq u.$$

(ii) *if for any $P \in \Theta_0$, F_P is continuous, we have for any realization x of X ,*

$$p(x) = \min\{\alpha \in [0, 1] : S(x) \geq \sup_{P \in \Theta_0} F_P^{-1}(1 - \alpha)\}.$$

If additionally Θ_0 is a singleton, $p(X) \sim U(0, 1)$ whenever $P \in \Theta_0$.

(iii) *if for any $P \in \Theta_0$, the variable $S(X)$ takes its values in a discrete set with probability 1, we have for any realization x of X ,*

$$p(x) = \min\{\alpha \in [0, 1] : S(x) > \sup_{P \in \Theta_0} F_P^{-1}(1 - \alpha)\}.$$

In particular, if $S(X)$ is an integer random variable, we have for any x such that $S(x) \in \mathbb{N}$,

$$p(x) = \min\{\alpha \in [0, 1] : S(x) \geq \sup_{P \in \Theta_0} F_P^{-1}(1 - \alpha) + 1\}.$$

A consequence is that the two classical definitions of a p -value are compatible in the following way.

Corollary A.2. Assume that there exists $Q \in \Theta_0$ such that for any $P \in \Theta_0$, for all $s \in \mathbb{R}$, $F_P(s) \geq F_Q(s)$. Let $p(X) = T_Q(S(X))$ and consider the families of tests $\{\phi_\alpha\}_{\alpha \in [0,1]}$ and $\{\phi'_\alpha\}_{\alpha \in [0,1]}$, where $\phi_\alpha(x) = \mathbf{1}\{S(x) \geq F_Q^{-1}(1 - \alpha)\}$ and $\phi'_\alpha(x) = \mathbf{1}\{S(x) > F_Q^{-1}(1 - \alpha)\}$. Then the following holds.

(i) if F_Q is continuous, the tests ϕ_α and ϕ'_α are of level α for all $\alpha \in [0, 1]$ and we have for any realization x of X ,

$$[p(x), 1] = \{\alpha \in [0, 1] : \phi_\alpha(x) = 1\}.$$

and for Q -almost every x ,

$$(p(x), 1] = \{\alpha \in [0, 1] : \phi'_\alpha(x) = 1\}.$$

(ii) if for $X \sim Q$ the variable $S(X)$ takes its values in a discrete set with probability 1, the test ϕ'_α is of level α while the test ϕ_α is not of level α , for all $\alpha \in [0, 1]$, and we have for any realization x of X ,

$$[p(x), 1] = \{\alpha \in [0, 1] : \phi'_\alpha(x) = 1\}.$$

In particular, we have both in the continuous and discrete case that for Q -almost every x ,

$$p(x) = \inf\{\alpha \in [0, 1] : \phi'_\alpha(x) = 1\}.$$

Proof. From Proposition A.1 (ii) and (iii), the only assertion to be proved is that for all $\alpha \in [0, 1]$, for Q -almost every x , $S(x) > F_Q^{-1}(1 - \alpha) \Leftrightarrow p(x) < \alpha$. Let us denote $\mathcal{Q} = \{F_Q^{-1}(1 - \alpha), \alpha \in [0, 1]\}$. Since F_Q is increasing on \mathcal{Q} , the desired relation is provided for $S(x) \in \mathcal{Q}$. We can conclude because $\mathbb{P}_{X \sim Q}(S(X) \in \mathcal{Q}) = 1$. □

Example A.3. To illustrate (i) and (iii) of Proposition A.1, let us consider the following simple discrete testing setting (coming from Example 3.3.2 in [37]). Let $H_0 : "P = P_0"$ where P_0 is the uniform distribution on $\{1, \dots, 10\}$ and consider the test statistic $S(X) = X$. We easily see that the p -value $T_{P_0}(X)$ is $p(X) = (11 - X)/10$. It satisfies $\mathbb{P}(p(X) \leq u) \leq u$, with equality iff u can be written under the form $i/10$ for some integer i , $1 \leq i \leq 10$. Furthermore, rejecting H_0 for $p(X) \leq \alpha$ is equivalent to reject H_0 whenever $X \geq k(\alpha)$ where $k(\alpha)$ is the unique integer satisfying $(11 - k(\alpha))/10 \leq \alpha < (12 - k(\alpha))/10$. We merely check that $k(\alpha) = F_{P_0}^{-1}(1 - \alpha) + 1$.

Finally, we provide a proof for Proposition A.1.

Proof. Let $\hat{F}_P(s) = \mathbb{P}_{X \sim P}(S(X) < s)$ and let us first state the following result: for any P , for any $\alpha \in [0, 1]$,

$$\{T_P(S(X)) \leq \alpha\} = \begin{cases} \{S(X) \geq F_P^{-1}(1 - \alpha)\} & \text{if } \hat{F}_P(F_P^{-1}(1 - \alpha)) = 1 - \alpha \\ \{S(X) > F_P^{-1}(1 - \alpha)\} & \text{otherwise} \end{cases}. \quad (41)$$

To establish (41), first note that $\{T_P(S(X)) \leq \alpha\} = \{\mathring{F}_P(S(X)) \geq 1 - \alpha\} \subset \{S(X) \geq F_P^{-1}(1 - \alpha)\}$, by definition of $F_P^{-1}(1 - \alpha)$. On the one hand, if $\mathring{F}_P(F_P^{-1}(1 - \alpha)) = 1 - \alpha$, we have $\{S(X) \geq F_P^{-1}(1 - \alpha)\} \subset \{\mathring{F}_P(S(X)) \geq \mathring{F}_P(F_P^{-1}(1 - \alpha))\} = \{\mathring{F}_P(S(X)) \geq 1 - \alpha\}$. On the other hand, if $\mathring{F}_P(F_P^{-1}(1 - \alpha)) < 1 - \alpha$, we have $\{\mathring{F}_P(S(X)) \geq 1 - \alpha\} \subset \{S(X) > F_P^{-1}(1 - \alpha)\}$ and $\{S(X) > F_P^{-1}(1 - \alpha)\} \subset \{\mathring{F}_P(S(X)) \geq F_P(F_P^{-1}(1 - \alpha))\} \subset \{\mathring{F}_P(S(X)) \geq 1 - \alpha\}$. This proves (41).

Let us now prove (i). We have for any $P \in \Theta_0$, $\mathbb{P}_{X \sim P}(p(X) \leq \alpha) \leq \mathbb{P}_{X \sim P}(T_P(S(X)) \leq \alpha)$. Next, applying (41), we have if $\mathring{F}_P(F_P^{-1}(1 - \alpha)) = 1 - \alpha$,

$$\mathbb{P}_{X \sim P}(p(X) \leq \alpha) \leq \mathbb{P}_{X \sim P}(S(X) \geq F_P^{-1}(1 - \alpha)) = 1 - \mathring{F}_P(F_P^{-1}(1 - \alpha)) = \alpha$$

and if $\mathring{F}_P(F_P^{-1}(1 - \alpha)) < 1 - \alpha$,

$$\mathbb{P}_{X \sim P}(p(X) \leq \alpha) \leq \mathbb{P}_{X \sim P}(S(X) > F_P^{-1}(1 - \alpha)) = 1 - F_P(F_P^{-1}(1 - \alpha)) \leq \alpha.$$

Assume now that for any $P \in \Theta_0$, F_P is continuous, and prove (ii). In this case, $\mathring{F}_P(F_P^{-1}(1 - \alpha)) = F_P(F_P^{-1}(1 - \alpha)) = 1 - \alpha$ for any $\alpha \in [0, 1]$, so that (41) provides that $\{T_P(S(X)) \leq \alpha\} = \{S(X) \geq F_P^{-1}(1 - \alpha)\}$. Hence, we obtain for any realization x of X ,

$$\begin{aligned} p(x) &= \min\{\alpha \in [0, 1] : \forall P \in \Theta_0, T_P(S(x)) \leq \alpha\} \\ &= \min\{\alpha \in [0, 1] : \forall P \in \Theta_0, S(x) \geq F_P^{-1}(1 - \alpha)\}, \end{aligned}$$

which leads to the desired result.

For (iii), the proof is similar by noting that $\mathring{F}_P(F_P^{-1}(1 - \alpha)) < 1 - \alpha$ in the case where the distribution of $S(X)$ has a discrete support under the null. \square

References

- [1] S. Arlot, G. Blanchard, and E. Roquain. Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.*, 38(1):51–82, 2010.
- [2] S. Arlot, G. Blanchard, and E. Roquain. Some nonasymptotic results on resampling in high dimension. II. Multiple tests. *Ann. Statist.*, 38(1):83–99, 2010.
- [3] Y. Baraud, S. Huet, and B. Laurent. Adaptive tests of linear hypotheses by model selection. *Ann. Statist.*, 31(1):225–251, 2003.
- [4] Y. Benjamini and R. Heller. False discovery rates for spatial signals. *J. Amer. Statist. Assoc.*, 102(480):1272–1281, 2007.
- [5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.
- [6] Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Behav. Educ. Statist.*, 25:60–83, 2000.
- [7] Y. Benjamini, A. M. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507, 2006.
- [8] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001.
- [9] M. A. Black. A note on the adaptive control of false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(2):297–304, 2004.
- [10] G. Blanchard. Contrôle non-asymptotique adaptatif du family-wise error rate en tests multiples. Talk at Journées Statistiques du Sud, Porquerolles, 2009.
- [11] G. Blanchard and F. Fleuret. Occam’s hammer. In *Learning theory*, volume 4539 of *Lecture Notes in Comput. Sci.*, pages 112–126. Springer, Berlin, 2007.

- [12] G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992, 2008.
- [13] G. Blanchard and E. Roquain. Adaptive false discovery rate control under independence and dependence. *J. Mach. Learn. Res.*, 10:2837–2871, 2009.
- [14] A. Celisse and S. Robin. A cross-validation based estimation of the proportion of true null hypotheses. *Journal of Statistical Planning and Inference*, 140(11):3132 – 3147, 2010.
- [15] Z. Chi and Z. Tan. Positive false discovery proportions: intrinsic bounds and adaptive control. *Statist. Sinica*, 18(3):837–860, 2008.
- [16] S. Delattre and E. Roquain. On the false discovery proportion convergence under gaussian equi-correlation. *Statistics & Probability Letters*, 81(1):111–115, 2011.
- [17] S. Dudoit and M. J. van der Laan. *Multiple testing procedures with applications to genomics*. Springer Series in Statistics. Springer, New York, 2008.
- [18] C. Durot and Y. Rozenholc. An adaptive test for zero mean. *Math. Methods Statist.*, 15(1):26–60, 2006.
- [19] B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23(1):1–22, 2008.
- [20] B. Efron. Correlated z -values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.*, 105(491):1042–1055, 2010.
- [21] A. Farcomeni. Some results on the control of the false discovery rate under dependence. *Scand. J. Statist.*, 34(2):275–297, 2007.
- [22] J. A. Ferreira and A. H. Zwinderman. On the Benjamini-Hochberg method. *Ann. Statist.*, 34(4):1827–1849, 2006.
- [23] H. Finner, R. Dickhaus, and M. Roters. On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.*, 37(2):596–618, 2009.
- [24] H. Finner, T. Dickhaus, and M. Roters. Dependency and false discovery rate: asymptotics. *Ann. Statist.*, 35(4):1432–1455, 2007.
- [25] R. A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.
- [26] Y. Gavrilov, Y. Benjamini, and S. K. Sarkar. An adaptive step-down procedure with proven FDR control under independence. *Ann. Statist.*, 37(2):619–629, 2009.
- [27] C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):499–517, 2002.
- [28] C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061, 2004.
- [29] J. Goeman and L. Finos. The inheritance procedure: multiple testing of tree-structured hypotheses. Technical report, Leiden University medical center, 2010.
- [30] J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *Ann. Statist.*, 38(6):3782–3810, 2010.
- [31] W. Guo and M. B. Rao. On control of the false discovery rate under no assumption of dependency. *Journal of Statistical Planning and Inference*, 138(10):3176–3188, 2008.
- [32] Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987.
- [33] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70, 1979.
- [34] K. In Kim, E. Roquain, and M. A. van de Wiel. Spatial clustering of array CGH features in combination with hierarchical multiple testing. *Stat. Appl. Genet. Mol. Biol.*, 9(1):Art. 40, 2010.
- [35] K. In Kim and M. van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9(1):114, 2008.
- [36] E. L. Lehmann and J. P. Romano. Generalizations of the familywise error rate. *Ann. Statist.*, 33:1138–1154, 2005.
- [37] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005.
- [38] Christopher J. M., Christopher G., R. C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, J. Schneider, and A. Moore. Controlling the false-discovery rate in astrophysical data analysis. *The Astronomical Journal*, 122(6):3492–3505, 2001.
- [39] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

- [40] N. Meinshausen, L. Meier, and P. Bühlmann. p-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, 104(488):1671–1681, 2009.
- [41] P. Neuvial. Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Stat.*, 2:1065–1110, 2008.
- [42] D. Pantazis, T. E. Nichols, S. Baillet, and R. M. Leahy. A comparison of random field theory and permutation methods for statistical analysis of MEG data. *NeuroImage*, 25:383–394, 2005.
- [43] J. P. Romano and A. M. Shaikh. Stepup procedures for control of generalizations of the familywise error rate. *Ann. Statist.*, 34(4):1850–1873, 2006.
- [44] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Statist. Assoc.*, 100(469):94–108, 2005.
- [45] J. P. Romano and M. Wolf. Control of generalized error rates in multiple testing. *Ann. Statist.*, 35(4):1378–1408, 2007.
- [46] J. P. Romano and M. Wolf. Balanced control of generalized error rates. *Ann. Statist.*, 38(1):598–633, 2010.
- [47] E. Roquain and M. van de Wiel. Optimal weighting for false discovery rate control. *Electron. J. Stat.*, 3:678–711, 2009.
- [48] E. Roquain and Fanny Villers. Exact calculations for false discovery proportion with application to least favorable configurations. *Ann. Statist.*, 39(1):584–612, 2011.
- [49] D. Rubin, S. Dudoit, and M. van der Laan. A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol.*, 5:Art. 19, 20 pp. (electronic), 2006.
- [50] S. K. Sarkar. Some results on false discovery rate in stepwise multiple testing procedures. *Ann. Statist.*, 30(1):239–257, 2002.
- [51] S. K. Sarkar. On methods controlling the false discovery rate. *Sankhya, Ser. A*, 70:135–168, 2008.
- [52] T. Schweder and E. Spjøtvoll. Plots of P-values to evaluate many tests simultaneously. *Biometrika*, 69(3):493–502, 1982.
- [53] P. Seeger. A note on a method for the analysis of significances en masse. *Technometrics*, 10(3):586–593, 1968.
- [54] V. G. Spokoiny. Adaptive hypothesis testing using wavelets. *Ann. Statist.*, 24(6):2477–2498, 1996.
- [55] J. D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498, 2002.
- [56] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):187–205, 2004.
- [57] M. A. van de Wiel, J. Berkhof, and W. N. van Wieringen. Testing the prediction error difference between 2 predictors. *Biostat*, 10(3):550–560, July 2009.
- [58] M. A. van de Wiel and W. N. van Wieringen. CGHregions: Dimension Reduction for Array CGH Data with Minimal Information Loss. *Cancer Inform*, 3:55–63, 2007.
- [59] M. J. van der Laan, M. D. Birkner, and A. E. Hubbard. Empirical Bayes and resampling based multiple testing procedure controlling tail probability of the proportion of false positives. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 29, 32 pp. (electronic), 2005.
- [60] N. Verzelen and F. Villers. Goodness-of-fit tests for high-dimensional Gaussian linear models. *Ann. Statist.*, 38(2):704–752, 2010.
- [61] L. Wasserman. *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004. A concise course in statistical inference.
- [62] L. Wasserman and K. Roeder. Weighted hypothesis testing. Technical report, Dept. of statistics, Carnegie Mellon University, 2006.
- [63] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing*. Wiley, 1993. Examples and Methods for P-Value Adjustment.