# Adaptive FDR control under independence and dependence

Gilles Blanchard, Etienne Roquain

▶ **To cite this version:**

Gilles Blanchard, Etienne Roquain. Adaptive FDR control under independence and dependence. 2008. hal-00159723v2

**HAL Id: hal-00159723**

**https://hal.science/hal-00159723v2**

Preprint submitted on 8 Apr 2008 (v2), last revised 17 Feb 2009 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

hal-00159723, version 2 - 8 Apr 2008

# Adaptive FDR control under independence and dependence

Gilles Blanchard, Étienne Roquain

April 8, 2008

**Abstract**

In the context of multiple hypotheses testing, the proportion $\pi_0$ of true null hypotheses among the hypotheses to test is a quantity that often plays a crucial role, although it is generally unknown. In order to obtain more powerful procedures, recent research has focused on finding ways to estimate this proportion and incorporate it in a meaningful way in multiple testing procedures, leading to so-called "adaptive" procedures. In this paper we focus on the issue of False Discovery Rate (FDR) control and we present new adaptive multiple testing procedures with control of the FDR, respectively under independence, positive dependencies (PRDS) or unspecified dependencies between the $p$-values. First, we present a new "one-stage" adaptive procedure and a new "two-stage" adaptive procedure that control the FDR in the independent context. We also give a review of other existing adaptive procedures that have provably controlled FDR in this context, and report extensive experimental results comparing these procedures and testing their robustness when the independence assumption is violated. Secondly, we propose adaptive versions of step-up procedures that have provably controlled FDR under positive dependencies and unspecified dependencies of the $p$-values, respectively. These are to our knowledge among the first theoretically founded adaptive multiple testing procedures that control the FDR when the $p$-values are not independent.

## 1 Introduction

Multiple testing is a topic coming from statistics that has generated growing attention in the recent years, spurred by an increasing number of application fields with a strong demand for powerful, large scale multiple testing procedures, like bioinformatics. For example, with microarray data, the goal is to detect which genes (among several ten of thousands) exhibit a significantly different level of expression in two different experimental conditions. Each gene represents a "hypothesis" to be tested in the statistical sense. The genes' expression levels fluctuate naturally (not to speak of other sources of fluctuation introduced by the experimental protocol), and, because they are so many genes to choose from, it is important to control precisely what can be deemed a significant observed difference. Generally it is assumed that the natural fluctuation distribution of a *single* gene is known and the problem is to take into account the number of genes involved (for more details, see for instance [1]).

In this work, we focus on building multiple testing procedures with a control of the false discovery rate (FDR). This quantity is defined as the expected proportion of type I errors, that is, the proportion of true null hypotheses among all the null hypotheses that have been rejected (i.e. declared as false) by the procedure. In their seminal work on this topic, Benjamini and Hochberg [2] proposed the celebrated *linear step-up* (LSU) procedure, that is

proved to control the FDR under independence between the $p$-values. Later, it was proved [3] that the LSU procedure still controls the FDR when the $p$-values have positive dependencies (or more precisely, a specific form of positive dependency called PRDS). Under unspecified dependencies, the same authors have shown that the FDR control still holds if the threshold collection of the LSU procedure is divided by a factor $1 + 1/2 + \cdots + 1/m$, where $m$ is the total number of null hypotheses to test. More recently, the latter result has been generalized [4], by showing that there is a family of step-up procedures (depending on the choice of a kind of prior distribution) that still control the FDR under unspecified dependencies between the $p$-values.

However, all these procedures, which are built in order to control the FDR at a level $\alpha$, can be showed to have actually their FDR upper bounded by $\pi_0\alpha$, where $\pi_0$ is the proportion of true null hypotheses. Therefore, when most of the hypotheses are false (i.e., $\pi_0$ is small), these procedures are inevitably conservative, since their actual FDR is much lower than the fixed target $\alpha$. In this context, the challenge of *adaptive control* of the FDR (see *e.g.* [5] and [6]) is to integrate an estimation of the unknown proportion $\pi_0$ in the threshold of the previous procedures and to prove that the FDR is still rigorously controlled by $\alpha$.

This adaptivity problem appears for instance when using hierarchical procedures (see eg [7]) which first selects some clusters of hypotheses that are likely to contain false nulls, and then apply a multiple testing procedure on the selected hyptheses. Since a large part of the true null hypotheses is expected to be false in the second step, an adaptive procedure is needed in order to keep the FDR close to the target level.

A popular way to estimate the proportion $\pi_0$ is the Storey estimator, initially introduced in [8], based on the suitably corrected proportion of observed $p$-values larger than a threshold $\lambda$ fixed by the user. This estimator can then be directly "plugged in" as a multiplicative level correction to the usual step-up procedures. It was recently proved [9; 10] that such a plug-in procedure (or a variation thereof) has provably controlled FDR under independence of the $p$-values.

A slightly different class of adaptive estimators consists of so-called *two-stage* procedures. In this approach, a first round of multiple hypothesis testing is performed using some fixed algorithm, then the results of this first round are used in order to tune the parameters of a second round in an adaptive way (this can generally be interpreted as using the output of the first stage to estimate $\pi_0$). Different procedures following this general approach have been proposed in [9; 11; 12].

Finally, some other works [13; 10; 14] have studied the question of adaptivity to the parameter $\pi_0$ from an *asymptotic* viewpoint. In this framework the more specific *random effects* model is assumed, where $p$-values are assumed independent, each hypothesis has a probability $\pi_0$ of being true, and all false null hypotheses share the same alternate distribution. The behavior of different procedures is then studied under ther limit where the number of tested hypotheses grows to infinity. One advantage of this approach and specific model is that it allows to derive quite precise results (see [15] for a precise study of limiting behaviors of central limit type under this model). However, we emphasize that in the present work our focus is decidedly on the nonasymptotic side, using finite samples and arbitrary alternate hypotheses.

The contributions of the present paper in this framework are the following. A first goal of the paper is to intoduce a number of novel adaptive procedures:

1. We introduce a simple new step-up procedure that is more powerful in general than the

standard LSU procedure, and provably controls the FDR under independence. This procedure is called one-stage adaptive, because the estimation of $\pi_0$ is performed implicitly.

2. Based on this, we then build a new two-stage adaptive procedure, which is more powerful in general than the procedure proposed in [9] while provably controlling the FDR under independence.

3. Under the assumption of positive or arbitrary dependence of the $p$-values, we introduce new two-stage adaptive versions of known step-up procedures (namely, of the LSU under positive dependences, and of the family of procedures described in [4] under unspecified dependencies). These adaptive versions provably control the FDR and result in an improvement of power over the non-adaptive versions in some situations (namely, when the number of hypotheses rejected in the first stage is large, typically more than 60%).

A second goal of this work is to present a review and comparative simulation study of several existing adaptive step-up procedures with provable FDR control (under independence). We present the theoretical FDR control as a consequence of a single general theorem for plug-in procedures, which was first established in [9]. Here, we present a self-contained proof of this result, that we think is more synthetic and therefore of independent interest, based on some tools introduced earlier by us in [16]. We compare these existing procedures to our new ones in extensive simulations following the simulation model and methodology used in [9].

The paper is organized as follows: in Section 2, we introduce the mathematical framework, and we recall the existing non-adaptive results in FDR control. In Section 3 we deal with the setup of independent $p$-values. We expose our new procedures and review the existing ones, and finally compare them in a simulation study. The case of positive dependent and arbitrarily dependent $p$-values is examined in Section 4 where we introduce our new adaptive procedures in this context. Section 5 contains proofs of the results. Some technical remarks and discussions of links to other work are gathered at the end of each relevant subsection, and can be skipped by the non-specialist reader.

# 2 Preliminaries

## 2.1 Multiple testing framework

In this paper, we will stick to the traditional statistical framework for multiple testing. Let $(\mathcal{X}, \mathfrak{X}, \mathbb{P})$ be a probability space; we want to infer a decision on $\mathbb{P}$ from an observation $x$ on $\mathcal{X}$ drawn from $\mathbb{P}$. Let $\mathcal{H}$ be a finite set of null hypotheses for $\mathbb{P}$, that is, each null hypothesis $h \in \mathcal{H}$ corresponds to some subset of distributions on $(\mathcal{X}, \mathfrak{X})$ and "$\mathbb{P}$ satisfies $h$" means that $\mathbb{P}$ belongs to this subset of distributions. The number of null hypotheses $|\mathcal{H}|$ is denoted by $m$. The underlying probability $\mathbb{P}$ being fixed, we denote $\mathcal{H}_0 = \{h \in \mathcal{H} | \mathbb{P} \text{ satisfies } h\}$ the set of the true null hypotheses and $m_0 = |\mathcal{H}_0|$ the number of true null hypotheses. We let also $\pi_0 := m_0/m$ the proportion of true null hypotheses. Since $\mathbb{P}$ is unknown, we remark that $\mathcal{H}_0$, $m_0$, and $\pi_0$ are unknown.

We suppose that there exists a set of *p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$, meaning that each $p_h : (\mathcal{X}, \mathfrak{X}) \mapsto [0,1]$ is a measurable function and that for each $h \in \mathcal{H}_0$, $p_h$ is bounded stochastically by a uniform distribution, that is,

$$\forall h \in \mathcal{H}_0 \qquad \forall t \in [0,1], \ \mathbb{P}(p_h \leq t) \leq t. \tag{1}$$

Typically, $p$-values are obtained from statistics that have a known distribution $P_0$ under the corresponding null hypothesis. In this case, if $F_0$ denotes the corresponding cumulative distribution function, applying $1-F_0$ to the observed statistic results in a random variable satisfying (1) in general. Here, we are however not concerned how these $p$-values are constructed and assume that they exist and are known.

### 2.2 Multiple testing procedure and errors

A *multiple testing procedure* is a measurable function

$$R : [0,1]^{\mathcal{H}} \mapsto \mathcal{P}(\mathcal{H}),$$

which takes as input a realization of the $p$-values and returns a subset of $\mathcal{H}$, corresponding to the rejected hypotheses. From an observation $x \in \mathcal{X}$, the procedure $R$ rejects the null hypotheses in the set $R(\mathbf{p}(x))$. To simplify the notations in what follows, we will often write $R$ instead of $R(\mathbf{p})$.

A multiple testing procedure $R$ can make two kinds of errors: a *type I error* occurs for $h$ when $h$ is true and is rejected by $R$, that is, $h \in \mathcal{H}_0 \cap R$. Following the Neyman-Pearson general philosophy for hypothesis testing, the primary concern is to control the number of type I errors of a testing procedure. Conversely, *a type II error* occurs for $h$ when $h$ is false and is not rejected by $R$, that is $h \in \mathcal{H}_0^c \cap R^c$.

The most traditional way to control type I error is to upper bound the "Family-wise error rate" (FWER), which is the probability that one or more true null hypotheses are wrongly rejected. However, procedures with a controlled FWER are very "cautious" not to make a single error, and thus reject only few hypotheses. This conservative way of measuring the type I error for multiple hypothesis testing can be a serious hinderance in practice, since it requires to collect large enough datasets so that significant evidence can be found under this strict error control criterion. More recently, a more liberal measure of type I errors has been introduced in multiple testing (see [2]): the *false discovery rate* (FDR), which is the averaged proportion of true null hypotheses in the set of all the rejected hypotheses:

**Definition 2.1 (False discovery rate).** The *false discovery rate* of a multiple testing procedure $R$ is given by

$$\mathrm{FDR}(R) := \mathbb{E}\left(\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{|R| > 0\}\right), \tag{2}$$

where $|.|$ is the cardinality function.

*Remark* 2.2. Throughout this paper we will use the following convention: whenever there is an indicator function inside an expectation, this has logical priority over any other factor appearing in the expectation. What we mean is that if other factors include expressions that may not be defined (such as the ratio $\frac{0}{0}$) outside of the set defined by the indicator, this is safely ignored. This results in more compact notations, such as in the above definition.

The goal, then, is to build procedures $R$ with a FDR upper bounded at a given level $\alpha$. Of course, if we choose $R = \emptyset$, meaning that $R$ rejects no hypotheses, $\mathrm{FDR}(R) = 0 \leq \alpha$ trivially. Therefore, it is desirable to build procedures $R$ satisfying $\mathrm{FDR}(R) \leq \alpha$ while at the same time having as small a type II error as possible. As a general rule, provided that $\mathrm{FDR}(R) \leq \alpha$, we therefore want to build procedures that reject as many hypotheses as possible. To this end we introduce the following notion: given two procedures $R$ and $R'$ that satisfy $\mathrm{FDR}(R) \leq \alpha$ and

$\mathrm{FDR}(R') \leq \alpha$, we say that $R$ is said *(uniformly) less conservative* than $R'$ if $R' \cap \mathcal{H}_0^c \subset R \cap \mathcal{H}_0^c$ pointwise. In particular, if $\mathrm{FDR}(R), \mathrm{FDR}(R') \leq \alpha$, $R$ is less conservative than $R'$ if $R \subset R'$ pointwise.

### 2.3 Step-up procedures, FDR control and adaptivity

In what follows, we sort the $p$-values in increasing order using the notation $p_{(1)} \leq \cdots \leq p_{(m)}$ and put $p_{(0)} = 0$. This order is of course itself random since it depends on the observation. We now define a widely used class of multiple testing procedures called *step-up procedures*.

**Definition 2.3 (Step-up procedure).** Let us fix $\alpha \in (0,1)$ and a positive nondecreasing function $\beta : \mathbb{R}^+ \mapsto \mathbb{R}^+$, called *shape function*. The *step-up procedure* of shape function $\beta$ (and at level $\alpha$) is defined as

$$R_\beta := \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}, \text{ where } k = \max\{0 \leq i \leq m \mid p_{(i)} \leq \alpha\beta(i)/m\}.$$

The function $\Delta(i) = \alpha\beta(i)/m$ is called the *threshold collection* of the procedure. In the particular case where the shape function $\beta$ is the identity function on $\mathbb{R}^+$, the procedure is called the *linear step-up procedure* (at level $\alpha$).

The linear step-up procedure (LSU) plays a prominent role in multiple testing under FDR control; it was the first procedure for which FDR control was proved and it is probably the most widely used procedure in this context. More precisely, when the $p$-values are assumed to be independent, the following theorem holds (the first part was proved in [2] whereas the second part was proved in [3; 17]):

**Theorem 2.4.** *Suppose that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent. Then the linear step-up procedure has FDR upper bounded by* $\pi_0\alpha$*, where* $\pi_0 = m_0/m$ *is the proportion of true null hypotheses. Moreover, if the p-values associated to true null hypotheses are exactly distributed like a uniform distribution, the linear step-up procedure has a FDR equal to* $\pi_0\alpha$ *.*

The authors [3] extended the previous result about FDR control of the linear step-up procedure to the case of $p$-values with a certain form of positive dependency called *positive regressive dependency from a subset* (PRDS). We skip a formal definition for now (we will get back to this topic in Section 4). In this particular dependency framework, the first part of the above theorem remains true.

However, when no particular assumptions are made on the dependencies between the $p$-values, it can be shown that the above FDR control is not generally true. This situation is called *unspecified* or *arbitrary* dependency. A modification of the LSU was first proposed in [3] which was proved to have a controlled FDR under arbitrary dependency. This result was extended in [4], where it is shown that there is a class of step-up procedures that control the FDR, as summed up in the following theorem:

**Theorem 2.5.** *Under unspecified dependencies between the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$*, consider* $\beta$ *a shape function of the form:*

$$\beta(r) = \int_0^r u d\nu(u), \tag{3}$$

*where* $\nu$ *is some fixed a priori probability distribution on* $(0, \infty)$*. Then the corresponding step-up procedure* $R_\beta$ *with threshold collection* $\Delta(i) = \alpha\beta(i)m$ *has FDR upper bounded by* $\alpha\pi_0$ *.*

In all of the above cases, the FDR is actually controlled at the level $\pi_0\alpha$ instead of the target $\alpha$. Hence, a direct corollary is that the step-up procedure $R_{\beta^\star}$ with $\beta^\star = \beta/\pi_0$ has FDR upper bounded $\alpha$ in either of the following situations:

- $\beta(x) = x$ when the $p$-values are independent or PRDS,
- the shape function $\beta$ is of the form (3) when the $p$-values have unspecified dependencies.

Note that, since $\pi_0 \leq 1$, the procedure $R_{\beta^\star}$ is always less conservative than $R_\beta$ (especially when $\pi_0$ is small). However, since $\pi_0$ is unknown, the procedure $R_{\beta^\star}$ cannot be derived from the observations only. We therefore will call the procedure $R_{\beta^\star}$ the *Oracle step-up procedure* of shape function $\beta$ (and at level $\alpha$).

Simply put, the role of adaptive step-up procedures is to mimic the latter oracle in order to obtain more powerful procedures. Adaptive procedures are often of the form as $R_{\beta G}$, where $G$ is some estimator of $\pi_0^{-1}$.

**Definition 2.6 (Plug-in adaptive step-up procedure).** Given a level $\alpha \in (0,1)$, a *shape function* $\beta$ and an estimator $G : [0,1]^{\mathcal{H}} \to (0,\infty)$ of the quantity $\pi_0^{-1}$, the *plug-in adaptive step-up procedure* of shape function $\beta$ and using estimator $G$ (at level $\alpha$) $R_{\alpha,\beta G}$, is defined as

$$R_{\beta G} = \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}, \text{ where } k = \max\{i \mid p_{(i)} \leq \alpha\beta(i)G(\mathbf{p})/m\}.$$

The (data-dependent) function $\Delta(i) = \alpha\beta(i)G(\mathbf{p})/m$ is called the *threshold collection* of the adaptive procedure. In the particular case where the shape function $\beta$ is the identity function on $\mathbb{R}^+$, the procedure is called an *adaptive linear step-up procedure* using estimator $G$ (and at level $\alpha$).

Following the previous definition, an adaptive plug-in procedure is composed of two different steps:

1. Estimate $\pi_0^{-1}$ with an estimator $G$.
2. Take the step-up procedure of shape function $\beta G$.

A subclass of plug-in adaptive procedures is formed by so-called *two-stage procedures*, when the estimator $G$ is actually based on a first, non-adaptive, multiple testing procedure. This can obviously be possibly iterated and lead to multi-stage procedures. The distinction between generic plug-in procedures and two-stage procedures is somewhat informal and generally meant only to provide some kind of nomenclature between different possible approaches.

The main theoretical task is to ensure that an adaptive procedure of this type still correctly controls the FDR. The mathematical difficulty obviously comes from the additional random variations of the estimator $G$ in the procedure.

# 3 Results under independence

In this section, we introduce two new procedures that adaptively control the FDR. The first one is one-stage and does not include an explicit estimator of $\pi_0^{-1}$, hence it is not a plug-in procedure. We then propose to use this as the first stage in a new two-stage procedure, which constitutes the second proposed method.

For clarity, we first introduce the new one-stage procedure; we then discuss several possible plug-in procedures, including our new proposition and several procedures proposed by other authors. FDR control for these various plug-in procedures can be studied using a general theoretical device introduced in [9] which we reproduce here with a self-contained and simpler proof. Finally, we close this section with extensive simulations to compare these different approaches.

### 3.1 A new adaptive one-stage step-up procedure

We present here our first main contribution, a "one-stage" adaptive step-up procedure. This means that the estimation step is implicitly included in the shape function $\beta$, and that this procedure is not of the plug-in type.

**Theorem 3.1.** *Suppose that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent and let* $\lambda \in (0,1)$ *be fixed. Then, the step-up procedure with threshold collection*

$$\Delta(i) = \min\left((1-\lambda)\frac{\alpha i}{m-i+1}, \lambda\right),$$

*has FDR upper bounded by* $\alpha$.

Below, we will make the choice $\lambda = \alpha$, leading to the threshold collection

$$\Delta(i) = \alpha \min\left((1-\alpha)\frac{i}{m-i+1}, 1\right). \tag{4}$$

For $i \leq (m+1)/2$, the threshold (4) is

$$\alpha\frac{(1-\alpha)i}{m-i+1},$$

and thus our approach differs from the threshold collection of the standard LSU procedure threshold by the factor $\frac{(1-\alpha)m}{m-i+1}$.

It is interesting to note that the correction factor $\frac{m}{m-i+1}$ appears in Holm's step-down procedure [18] for FWER control. The latter is a well-known improvement of Bonferroni's procedure (which corresponds to the fixed threshold $\alpha/m$), taking into account the proportion of true nulls, and defined as the step-down procedure with threshold collection $\alpha/(m-i+1)$. Here we therefore prove that this correction is suitable as well in the framework of FDR control.

As Figure 1 illustrates, our procedure is generally less conservative than the (non-adaptive) linear step-up procedure (LSU). Precisely, the new procedure can only be more conservative than the LSU procedure in the marginal case where the factor $\frac{(1-\alpha)m}{m-i+1}$ is smaller than one. This can happen only when the proportion of null hypotheses rejected by the LSU procedure is positive but less than $1/m+\alpha$, and even in this zone the ratio of the two threshold collections is never less than $(1-\alpha)$.

It is nevertheless of interest to determine if this unfavorable situation can occur. To investigate this issue, the next lemma considers a specific "Gaussian random effects" model (which is relatively standard in the multiple testing literature, see e.g. [13]) and gives a quantitative answer from an asymptotical point of view (when the number of tested hypotheses grows to infinity). In the random effect model, hypotheses are assumed to be randomly true or false with probability $\pi_0$, and the false null hypotheses share a common distribution $P_1$. Globally, the $p$-values then are i.i.d. drawn according to the mixture distribution $\pi_0 U[0,1] + (1-\pi_0)P_1$.

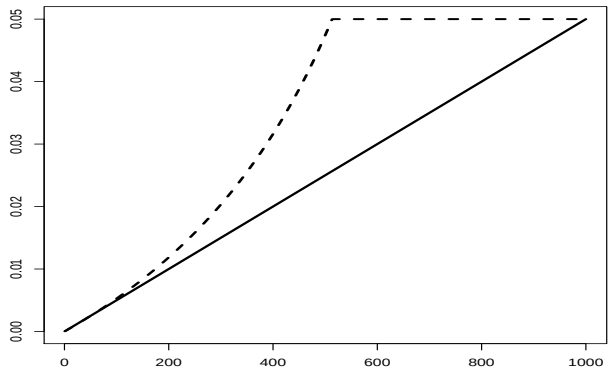


Figure 1: For $m = 1000$ null hypotheses and $\alpha = 0.05$: the new threshold collection (4) (dashed line) and the threshold collection of the linear procedure $\Delta(i) = \alpha i/m$ (solid line).

**Lemma 3.2.** *Consider the random effects model where the p-values are i.i.d. with common cumulative distribution function $t \mapsto \pi_0 t + (1-\pi_0)F(t)$. Assume the true null hypotheses are standard Gaussian with zero mean and the alternative hypotheses are standard Gaussian with mean $\mu > 0$. In this case $F(t) = \overline{\Phi}(\overline{\Phi}^{-1}(t) - \mu)$, where $\overline{\Phi}$ is the standard Gaussian upper tail function. Assuming $\pi_0 < (1+\alpha)^{-1}$, define*

$$\mu^\star = \overline{\Phi}^{-1}(\alpha^2) - \overline{\Phi}^{-1}\left(\frac{\alpha^{-1} - \pi_0}{1-\pi_0}\alpha^2\right).$$

*Then if $\mu > \mu^*$, the probability that the LSU rejects a proportion of null hypotheses less than $1/m + \alpha$ tends to 0 as $m$ tends to infinity. On the other hand, if $\pi_0 > (1+\alpha)^{-1}$, or $\mu < \mu^*$, then this probability tends to one.*

For instance, taking in the above lemma the values $\pi_0 = 0.5$ and $\alpha = 0.05$, results in the critical value $\mu^\star \simeq 1.51$. This lemma delineates clearly in a particular case in which situation we can expect an improvement from the adaptive procedure over the standard LSU.

*Remark* 3.3. It was recently pointed out to us that a one-stage adaptive procedure that is very similar to ours is proposed in the forthcoming paper [14]. The present work was developed independently. More precisely, the procedure proposed in [14] starts with some heuristic motivations leading to the threshold collection $t(i) = \frac{\alpha i}{m-(1-\alpha)i}$. However, this threshold collection, as is, does not control the FDR (since the corresponding step-up procedure would always reject all the hypotheses), and several suitable modifications are proposed in [14], the simplest one being $t'_\eta(i) = \min(t(i), \eta i/m)$. The theoretical FDR control proved in [14] is studied specifically under the "random effect" model, and only asymptotically as the number of hypotheses grows to infinity. In that framework, asymptotical control at level $\alpha$ holds for any $\eta < 1$ (but it is intuitively clear that the convergence is slower as $\eta$ gets close to 1). While the threshold collection proposed by us here is arguably slightly more conservative, Theorem 3.1 shows that it has provably controlled FDR under any distribution for the false hypotheses, and for any fixed finite number of hypotheses. Furthermore, we can use it in a 2-step procedure as will be argued in the next section.

If $r$ denotes the final number of rejections of the new one-stage procedure, we can interpret the ratio $\frac{(1-\lambda)m}{m-r+1}$ between the adaptive threshold and the LSU threshold at the same point as an *a posteriori* estimate for $\pi_0^{-1}$. In the next section we propose to use this quantity in a plug-in, 2-stage adaptive procedure.

### 3.2 Adaptive plug-in methods with provable FDR control

In this section, we consider different adaptive step-up procedures of the plug-in type, i.e. based on an explicit estimator of $\pi_0^{-1}$. We first review a general method proposed in [9] in order to derive FDR control for such plug-in procedures. A self-contained proof of this result is proposed, which is in our opinion more compact and synthetic than the original one. It relies on an approach and tools that were introduced in [16]. Based on this, we review the different plug-in estimators considered in [9] and add a new one to the lot, based on the one-stage adaptive procedure of the previous section.

Let us first introduce the following notations: for each $h \in \mathcal{H}$, we denote by $\mathbf{p}_{-h}$ the collection of $p$-values $\mathbf{p}$ restricted to $\mathcal{H} \setminus \{h\}$, that is, $\mathbf{p}_{-h} = (p_{h'}, h' \neq h)$. We also denote $\mathbf{p}_{0,h} = (\mathbf{p}_{-h}, 0)$ the collection $\mathbf{p}$ where $p_h$ has been replaced by 0.

**Theorem 3.4 (Benjamini, Krieger, Yekutieli).** *Suppose that the family $p$-values $\mathbf{p} = (p_h, h \in \mathcal{H})$ is independent and consider a measurable function $G : [0,1]^{\mathcal{H}} \to (0,\infty)$ coordinate-wise non-increasing, such that for each $h \in \mathcal{H}_0$,*

$$\mathbb{E}G(\mathbf{p}_{0,h}) \leq \pi_0^{-1}. \tag{5}$$

*Then, the adaptive linear step-up procedure $R$ of threshold collection $\Delta(i) = \alpha i G(\mathbf{p})/m$ has a FDR smaller than $\alpha$.*

We will apply the above result to the following estimators, depending on a fixed parameter $\lambda \in (0,1)$ or $k_0 \in \{1,\dots,m\}$:

$$\text{[Storey-}\lambda\text{]} \quad G_1(\mathbf{p}) = \frac{(1-\lambda)m}{\sum_{h\in\mathcal{H}} \mathbf{1}\{p_h > \lambda\} + 1};$$

$$\text{[Quant-}\frac{k_0}{m}\text{]} \quad G_2(\mathbf{p}) = \frac{(1-p_{(k_0)})m}{m-k_0+1};$$

$$\text{[BKY06-}\lambda\text{]} \quad G_3(\mathbf{p}) = \frac{(1-\lambda)m}{m-|R_0(\mathbf{p})|+1}, \text{ where } R_0 \text{ is the standard LSU at level } \lambda;$$

$$\text{[BR07-}\lambda\text{]} \quad G_4(\mathbf{p}) = \frac{(1-\lambda)m}{m-|R_0'(\mathbf{p})|+1}, \text{ where } R_0' \text{ is the new adaptive step-up of Theorem 3.1.}$$

Estimator $G_1$ is usally called *modified Storey's estimator* and was initially introduced by [8] from an heuristics on the $p$-values histogram (originally without the "+1" in the numerator, hence the name "modified"). Its intuitive justification is as follows: the set $S_\lambda$ of $p$-values larger than the threshold $\lambda$ contains on average at least a proportion $(1-\lambda)$ of the true null hypotheses. Hence, a natural estimator of $\pi_0^{-1}$ is $(1-\lambda)m/|S_\lambda \cap \mathcal{H}_0| \leq (1-\lambda)m/|S_\lambda| \simeq G_1(\mathbf{p})$. Therefore, we expect that Storey's estimator is generally an overestimate of $\pi_0^{-1}$. A standard choice is $\lambda = 1/2$ (as in the SAM software [19]). FDR control for the corresponding plug-in step-up procedure was proved asymptotically in [10] under the random effects model (actually, for the modification $\widetilde{G_1} = \min(G_1, \lambda)$); and in [9] in general under independence of the $p$-values.

Estimator $G_2$ was introduced in [5] and [20], from a slope heuristics on the $p$-values c.d.f. Roughly speaking, $G_2$ appears as a Storey's estimator with the data-dependent choice $\lambda = p_{(k)}$, and can therefore be interpreted as the quantile version of the Storey estimator. A classical value for $k$ is $\lfloor m/2 \rfloor$, resulting in the so-called median adaptive LSU (see [9]).

Estimator $G_3$ was introduced in [9] for the particular choice $\lambda = \alpha/(1+\alpha)$. More precisely, a slightly less conservative version, without the "+1" in the denominator, was used in [9]. We forget about this refinement here, noting that it results only in a very slight improvement.

Finally, the estimator $G_4$ is new and follows exactly the same philosophy as $G_3$, that is, uses a step-up procedure as a first stage in order to estimate $\pi_0^{-1}$, but this time based on our adaptive one-stage step-up procedure introduced in the previous section, rather than the standard LSU. Note that since $R_0'$ is less conservative than $R_0$ (except in marginal cases), we generally have $G_2 \leq G_3$ pointwise and our estimator improves the one of [9].

Note that the estimators $G_i$, $i = 1, 2, 3, 4$ are not necessarily larger than 1, and to this extent can in some unfavorable cases result in the final procedure being actually more conservative than the standard LSU. This will mainly happen if $\pi_0$ is close to 1; if such a situation is anticipated it is more appropriate to use the regular non-adaptive LSU.

These different estimators all satisfy condition (5), and we thus obtain the following corollary:

**Corollary 3.5.** *Assume that the p-values of* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *are independent. The plug-in adaptive linear step-up procedure at level* $\alpha$ *using any of the estimators* $G_1$ *to* $G_4$ *has a FDR smaller than* $\alpha$ *.*

The above result for $G_1$, $G_2$ (for $\lambda = \alpha/(1+\alpha)$) and $G_3$ was proved in [9]. For completeness, we reproduce shortly the corresponding arguments in the appendix.

*Remark* 3.6. The result proved in [9] is actually slightly sharper than Theorem 3.4. Namely, if $G(\cdot)$ is moreover supposed to be coordinate-wise left-continuous, it is possible to prove that Theorem 3.4 still holds when the condition (5) is replaced by the slightly weaker condition:

$$\mathbb{E}G(\widetilde{\mathbf{p}}_h) \leq \pi_0^{-1}, \tag{6}$$

where for each $h \in \mathcal{H}_0$, $\widetilde{\mathbf{p}}_h = (\mathbf{p}_{-h}, \widetilde{p}_h(\mathbf{p}_{-h}))$ is the collection of $p$-values $\mathbf{p}$ where $p_h$ has been replaced by $\widetilde{p}_h(\mathbf{p}_{-h}) = \max\left\{p \in [0,1] \;\middle|\; p \leq \alpha\pi(h)|R(\mathbf{p}_{-h}, p)|G(\mathbf{p}_{-h}, p)\right\}$. This improvement then permits to get rid of the "+1" in the denominator of $G_3$. Here, we opted for simplicity and a more straightforward statement, noting that this improvement is not crucial.

*Remark* 3.7. The one-stage step-up procedure of [14] (see previously Remark 3.3) — for which there is no result proving non-asymptotic FDR control up to our knowledge — can also be interpreted intuitively as an adaptive version of the LSU using estimator $G_2$, where the choice of parameter $k_0$ is data-dependent. Namely, assume we reject at least $i$ null hypotheses whenever $p_{(i)}$ is lower than the standard LSU threshold times the estimator $G_2(k_0)$, using parameter $k_0 = i$. This corresponds to the inequality $p_{(i)} \leq \frac{(1-p_{(i)})k}{m-i+1}$, which, solved in $p_{(i)}$, gives the threshold collection of [14]. Remember from Remark 3.3 that this threshold collection must actually be modified in order to be useful, since it otherwise always lead to reject all hypotheses. The modification mentioned earlier consists in capping the estimated $\pi_0^{-1}$ at a leverl $\eta$, i.e. using $\min(\eta, G_2)$ instead of $G_2$ in the above reasoning.

### 3.3 Simulation study

How can we compare the different adaptive procedures defined above? For a fixed $\lambda$, we have pointwise $G_1 \geq G_4 \geq G_3$ which shows that the adaptive procedure obtained using [Storey-$\lambda$] is always less conservative than the one derived from [BR08-$\lambda$], itself less conservative that the one using [BKY06-$\lambda$] (except in the marginal cases where the one-stage adaptive procedure is more conservative than the standard step-up procedure, delineated earlier). It would therefore appear that one should always choose [Storey-$\lambda$] and disregard the other ones. However, a important point made by [9] for introducing $G_3$ as a better alternative to the (already known earlier) $G_1$ is that, on simulations with positively dependent test statistics, the plug-in procedure using $G_1$ with $\lambda = 1/2$ had very poor control of the FDR, while the FDR was still controlled for the plug-in procedure based on $G_3$. While the positively dependent case is not covered by the theory, it is of course very important to ensure that a multiple testing procedure is sufficiently robust in practice so that the FDR does not vary too much in this situation.

In order to assess the quality of our new procedures, we compare here the different methods on a simulation study following the setting used by [9]. Let $X_i = \mu_i + \varepsilon_i$, for $i, 1 \leq i \leq m$, where $\varepsilon$ is a $\mathbb{R}^m$-valued centred Gaussian random vector such that $\mathbb{E}(\varepsilon_i^2) = 1$ and for $i \neq j$, $\mathbb{E}(\varepsilon_i \varepsilon_j) = \rho$, where $\rho \in [0, 1]$ is a correlation parameter. Thus, when $\rho = 0$ the $X_i$'s are independent, whereas when $\rho > 0$ the $X_i$'s are positively correlated (with a constant pairwise correlation). For instance, the $\varepsilon_i$'s can be constructed by taking $\varepsilon_i := \sqrt{\rho} U + \sqrt{1-\rho} Z_i$, where $Z_i$, $1 \leq i \leq m$ and $U$ are all i.i.d $\sim \mathcal{N}(0, 1)$.

Considering the one-sided null hypotheses $h_i$ : "$\mu_i \leq 0$" against the alternatives "$\mu_i > 0$" for $1 \leq i \leq m$, we define the $p$-values $p_i = \overline{\Phi}(X_i)$, for $1 \leq i \leq m$, where $\overline{\Phi}$ is the standard Gaussian distribution tail. We choose a common mean $\bar{\mu}$ for all false hypotheses, that is, for $i, 1 \leq i \leq m_0$, $\mu_i = 0$ and for $i, m_0 + 1 \leq i \leq m$, $\mu_i = \bar{\mu}$ ; the $p$-values corresponding to the null means follow exactly a uniform distribution.

We compare the following step-up multiple testing procedures:

- [*LSU Oracle*] the procedure with the threshold collection $\Delta(i) = \alpha i / m_0$.
- [*Storey-$\lambda$*] the plug-in procedure corresponding to $G_1$ in Corollary 3.5. A standard choice for $\lambda$ is $1/2$. Because of the relationship of $G_3, G_4$ to [Storey-$\alpha$], we also include $\lambda = \alpha$ .
- [*BKY06-$\alpha$*] The two-stage procedure corresponding to $G_3$ in Corollary 3.5 with parameter $\lambda = \alpha$ .
- [*BR08-1S-$\alpha$*] The new one-stage adaptive procedure of Theorem 3.1, with parameter $\lambda = \alpha$ .
- [*BR08-2S-$\alpha$*] The new two-stage adaptive procedure using the previous item in the first stage, corresponding to estimator $G_4$ in Corollary 3.5.
- [*FDR08-$\frac{1}{2}$*] The one-stage procedure proposed in [14] and described in Remark 3.3, with $\eta = \frac{1}{2}$ .
- [*Median LSU*] The plug-in procedure corresponding to estimator $G_2$ in Corollary 3.5, for $\frac{k_0}{m} = \frac{1}{2}$ .

Note that the procedure studied in [9] is actually [BKY06-$\alpha/(1+\alpha)$] in our notation. This means that the procedure used in our simulations is not exactly the same as in [9], but it is very close.

The three most important parameters in the simulation are the correlation coefficient $\rho$, the proportion of true null hypotheses $\pi_0$, and the alternative mean $\bar{\mu}$ which represents the signal-to-noise ration, or how easy it is to distinguish alternative hypotheses. We present in Figures 2, 3, and 4 results of the simulations for one varying parameter ($\pi_0$, $\bar{\mu}$ and $\rho$, respectively), the others being being kept fixed. Reported are, for the different methods: the average FDR, and the average power relative to the reference [LSU-Oracle]. The absolute power is defined as the average of false null hypotheses rejected, and the relative power as the mean of the number of true rejections of the procedure divided by the number of true rejections of [LSU-Oracle]. Each point is an average of $10^5$ simulations, with fixed paramaters $m = 100$ and $\alpha = 5\%$.

### 3.3.1 Under independence ($\rho = 0$)

Remember that under independence of the $p$-values, the *LSU* procedure has a FDR equal to $\alpha\pi_0$ and that the *LSU Oracle* procedure has a FDR equal to $\alpha$ (provided that $\alpha \leq \pi_0$). The other procedures have their FDR upper bounded by $\alpha$ (in an asymptotical sense only for [FDR08-$\frac{1}{2}$]).

The situation where the $p$-values are independent corresponds to the first row of Figures 2 and 3 and the leftmost point of each graph in Figure 4. It appears that in the indepedent case, the following procedures can be consistently ordered in terms of (relative) power over the range of parameters studied here:

$$[\textit{Storey-1/2}] \succ [\textit{Storey-}\alpha] \succ [\textit{BR08-2S-}\alpha] \succ [\textit{BKY06-}\alpha],$$

the symbol "$\succ$" meaning "is (uniformly over our experiments) more powerful than".

Next. the procedures [median-LSU] and [FDR08-$\frac{1}{2}$] appear both consistently less powerful than [Storey-$\frac{1}{2}$], and [FDR08-$\frac{1}{2}$] is additionally also consistently less powerful than [Storey-$\alpha$]. Their relation to the remaining procedures depends on the parameters; both [median-LSU] and [FDR08-$\frac{1}{2}$] appear to be more powerful than the remaining procedures when $\pi_0 > \frac{1}{2}$, and less efficient otherwise. We note that [median-LSU] also appears to perform better when $\bar{\mu}$ is low (i.e. the alternative hypotheses are harder to distinguish).

Finally, concerning our one-stage procedure [BR08-1S-$\alpha$], we note that it appears to be indistinguishable from its two-stage counterpart [BR08-2S-$\alpha$] when $\pi_0 > \frac{1}{2}$, and significantly less powerful otherwise. This also corresponds to our expectations, since in the situation $\pi_0 < \frac{1}{2}$, there is a much higher likelihood that more than 50% hypotheses are rejected, in which case our one-stage threshold family hits its "cap" at level $\alpha$ (see e.g. Fig. 1). This is precisely to improve on this situation that we introduced the 2-stage procedure, and we see that does in fact improve substantially the 1-stage version in that specific region.

The fact that [Storey-$\frac{1}{2}$] is uniformly more powerful than the other procedures in the independent case corroborates the simulations reported in [9]. Generally speaking, under independence we obtain a less biased estimate for $\pi_0^{-1}$ when considering Storey's estimator based on a "high" threshold like $\lambda = \frac{1}{2}$. Namely, higher $p$-values are less likely to be "contaminated" by false null hypotheses; conversely, if we take a lower threshold $\lambda$, there will be more false null hypotheses included in the set of $p$-values larger than $\lambda$, leading to a pessimistic bias in the estimation of $\pi_0^{-1}$. This qualitative reasoning is also consistent with the observed

behavior of [median-LSU], since the set of $p$-values larger than the median is much more likely to be "contaminated" when $\pi_0 < \frac{1}{2}$.

However, the problem with [Storey-$\frac{1}{2}$] is that the corresponding estimation of $\pi_0^{-1}$ exhibits much more variability than its competitors when there is a substantial correlation between the $p$-values. As a consequence it is a very fragile procedure. This phenomenon was already pinpointed in [9] and we study it next.

### 3.3.2 Under positive dependencies ($\rho > 0$)

Under positive dependencies, remember that it is known theoretically from [3] that the FDR of the procedure *LSU* (resp. *LSU Oracle*) is still bounded by $\alpha\pi_0$ (resp. $\alpha$), but without equality. However, we do not know from a theoretical point of view if the adaptive procedures have their FDR upper bounded by $\alpha$. In fact, it was pointed out in an other work reporting simulations on adaptive procedures [12], that one crucial point for these seems to be the variability of estimate of $\pi_0^{-1}$. Estimates of this quantity that are not robust with respect to positive dependence will result in failures for the corresponding multiple testing procedure.

The situation where the $p$-values are positively dependent corresponds to the second and third rows ($\rho = 0.2, 0.5$, respectively) of Figures 2 and 3 and to all the graphs of Figure 4.

The most striking fact is that [Storey-$\frac{1}{2}$] does not control the FDR at the desired level any longer under positive dependencies, an can even be off by quite a large factor. This is in accordance with the experimental findings in [9]. Therefore, although this procedure was the favorite in the independent case, it turns out to be not robust, which is very undesirable for practical use where it is generally impossible to guarantee that the $p$-values are independent. The procedure [median-LSU] appears to have higher power than the remaining ones in the situations studied in Figure 3, especially with a low signal-to-noise ratio. Unfortunately, other situations appearing in Figures 2 and 4 show that [median-LSU] can exhibit a very poor FDR control in some parameter regions. Furthermore, the behavior of this procedure appears somewhat erratic and complicated to interpret accross parameter values; in particular it seems difficult to characterize qualitatively and with some insight the regions where it has controlled FDR. Under these findings our conclusion is that [median-LSU] is also not robust enough in general to be reliable.

The other remaining procedures seem to still have a controlled FDR, or at least to be very close to the FDR target level. For these it seems that the qualitative conclusions concerning power comparison found in the independent case remain true. To sum up:

- the best overall procedure seems to be [Storey-$\alpha$]: its FDR seems to be under or only slightly over the target level in all situations, and it exhibits globally a power superior to other procedures.
- then come in order of power, our 2-stage procedure [BR08-2S-$\alpha$], then [BKY06-$\alpha$].
- like in the dependent case, [FDR08-$\frac{1}{2}$] ranks second when $\pi_0 > \frac{1}{2}$ but tends to perform noticeably poorer if $\pi_0$ gets smaller.

To conclude, a practical recommendation that we draw from these experiments is that for practical use, we recommend [Storey-$\alpha$], [BR08-2S-$\alpha$] or [FDR08-$\frac{1}{2}$] which all exhibit good robustness to dependence for FDR control as well as comparatively good power. The fact that [Storey-$\alpha$] performs so well and seems to hold the favorite position has up to our knowledge

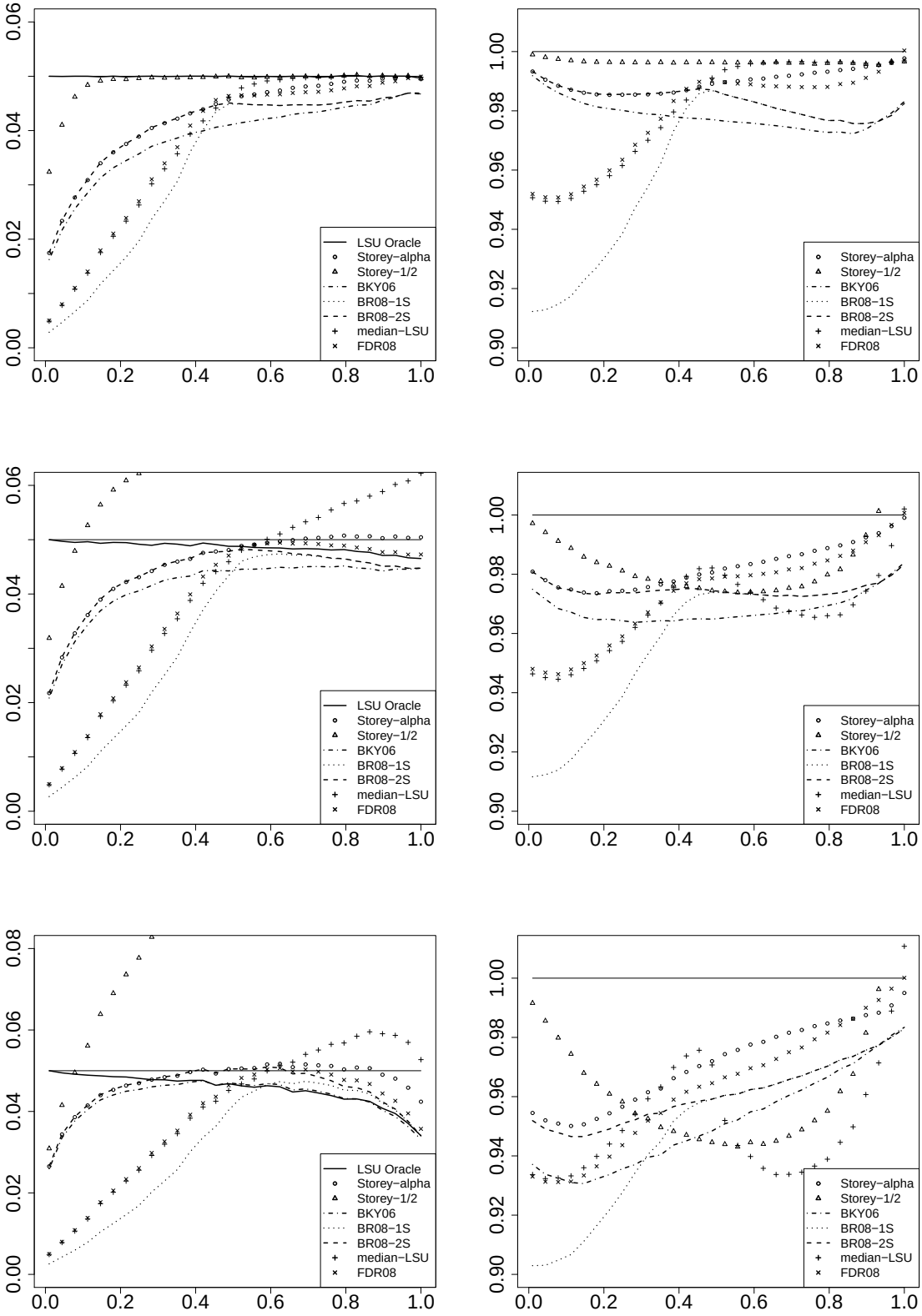


Figure 2: FDR (left column) and power relative to [LSU-Oracle] (right column) as a function of the true proportion $\pi_0$ of null hypotheses. Target FDR is $\alpha = 5\%$, total number of hypotheses $m = 100$. The mean for the alternatives is $\bar{\mu} = 3$. From top to bottom: pairwise correlation coefficient $\rho \in \{0, 0.2, 0.5\}$.

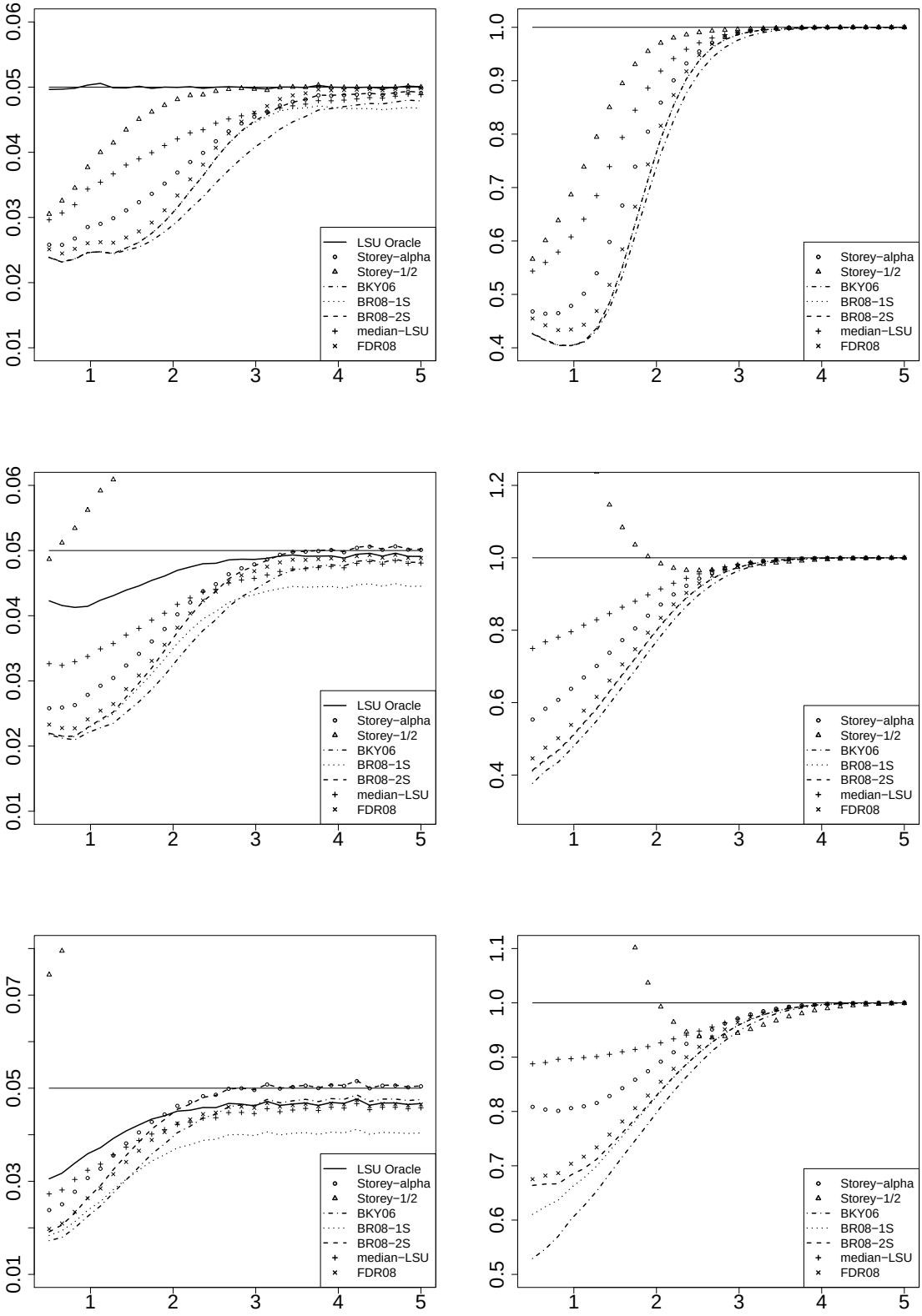


Figure 3: FDR (left column) and power relative to [LSU-Oracle] (right column) as a function of the alternative hypothesis mean $\mu$. Target FDR is $\alpha = 5\%$, total number of hypotheses $m = 100$. The proportion of true null hypotheses is $\pi_0 = 0.5$. From top to bottom: pairwise correlation coefficient $\rho \in \{0, 0.2, 0.5\}$.

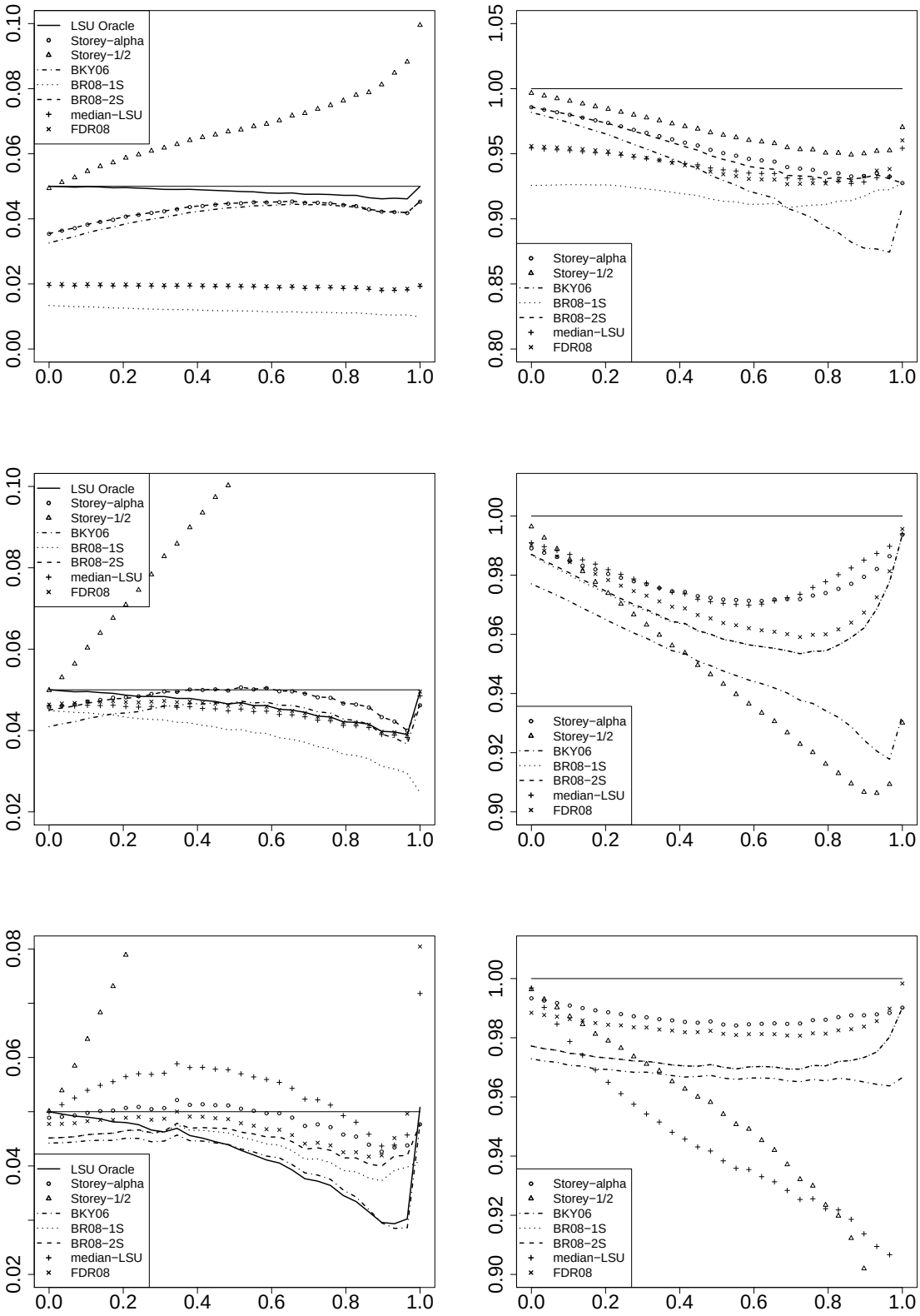


Figure 4: FDR (left column) and power relative to [LSU-Oracle] (right column) as a function of the pairwise correlation coefficient $\rho$.Target FDR is $\alpha = 5\%$, total number of hypotheses $m = 100$. The mean for the alternatives is $\bar{\mu} = 3$. From top to bottom: proportion of true null hypotheses $\pi_0 \in \{0.2, 0.5, 0.8\}$.

no been reported before (it was not included in the simulations in [9]) and came somewhat as a surprise to us.

*Remark* 3.8. The fact that [FDR08-$\frac{1}{2}$] performs sub-optimally for $\pi_0 < \frac{1}{2}$ appears to be strongly linked to the choice of parameter $\eta = \frac{1}{2}$. Namely, the implicit estimator of $\pi_0^{-1}$ in the procedure is capped at $\eta$ (see Remark 3.7). On the other hand, choosing a higher value for $\eta$ will necessarily yield a more variable and less reliable procedure (since the limit case $\eta = 1$ will result in all hypotheses being always rejected, see Remark 3.3). Here we chose the value $\frac{1}{2}$ as a reasonable default tradeoff.

*Remark* 3.9. Another 2-stage adaptive procedure was introduced in [11], which is very similar to a plug-in procedure using [Storey-$\lambda$]. In fact, in the experiments presented in [11], the two procedures are almost equivalent, corresponding to $\lambda = 0.995$. We decided not to include this additional procedure in our simulations to avoid overloading the plots. Qualitatively, we observed that the procedures of [11] or [Storey-0.995] are very similar in behavior to [Storey-$\frac{1}{2}$]: very performant in the independent case but very fragile with respect to deviations from independence.

# 4 Results under dependence

In this section, we consider from a theoretical point of view the problem of constructing multiple testing procedures that are adaptive to $\pi_0$ under dependency conditions of the $p$-values. The derivation of adaptive procedures that have provably controlled FDR under dependencies appears to have been only studied scarcely (see [11] and [12]). Here, we here propose to use a 2-stage procedure where the first stage is a multiple testing with either controlled FWER or controlled FDR. The first option is relatively straightfoward and is intended as a reference. In the second case, we use Markov's inequality to estimate $\pi_0^{-1}$. Since Markov's inequality is general but not extremely precise, the resulting procedures are obviously quite conservative and are arguably of a limited practical interest. However, we will show that they still provide an improvement, in a certain regime, with respect to (non-adaptive) LSU procedure in the PRDS case and with respect to the family of (non-adaptive) procedures proposed in Theorem 2.5 in the arbitrary dependences case.

For the purposes of this section, we first recall the formal definition for PRDS dependence from [3]:

**Definition 4.1.** Remember that a set $D \subset [0,1]^{\mathcal{H}}$ is said to be *non-decreasing* if for all $x, y \in [0,1]^{\mathcal{H}}$, $x \leq y$ coordinate-wise and $x \in D$ implies $y \in D$. Then, the $p$-values $\mathbf{p} = (p_h, h \in \mathcal{H})$ are said *positively regressively dependent on each one from* $\mathcal{H}_0$ (PRDS on $\mathcal{H}_0$ in short) if for any non-decreasing measurable set $D \subset [0,1]^{\mathcal{H}}$ and for all $h \in \mathcal{H}_0$, $u \in [0,1] \mapsto \mathbb{P}(\mathbf{p} \in D | p_h = u)$ is non-decreasing.

Remember that it was proved in [3] that the LSU still has controlled FDR at level $\pi_0 \alpha$ (i.e., Theorem 2.4 still holds under the PRDS assumption). On the other hand, recall Theorem 2.5 for a threshold collection resulting in controlled FDR at the same level under totally arbitrary dependences.

Our first result concerns a two-stage procedure where the first stage $R_0$ is a multiple testing with controlled FWER, and where we (over)estimate $m_0$ via the straightforward estimator $(m - |R_0|)$. This should be considered as a form of baseline reference for this type of two-stage procedure.

**Theorem 4.2.** *Assume $R_0$ is a multiple testing procedure with FWER controlled at level $\alpha_0$, that is, $\mathbb{P}[R_0 \cap \mathcal{H}_0 \neq \emptyset] \leq \alpha_0$, and such that number of rejected hypotheses $|R_0|$ is nonincreasing as a function of each p-value. Then the adaptive step-up procedure $R$ with data-dependent threshold collection $\Delta(i) = \alpha_1(m - |R_0|)^{-1}\beta(i)$ has FDR controlled at level $\alpha_0 + \alpha_1$ in either of the following dependence situations:*

- *the p-values $(p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$ and the shape function is the identity function.*
- *the p-values have unspecified dependencies and $\beta$ is a shape function of the form (3).*

Here it is clear that the price for adaptivity is a certain loss in FDR control for being able to use the information of the first stage. If we choose $\alpha_0 = \alpha_1 = \alpha/2$, then this procedure will outperform the non-adaptive $R_\beta$ only if there are more than $50\%$, rejected hypotheses in the first stage. Only if it is expected that this situation will occur does it make sense to employ this procedure, since it will otherwise perform worse than the non-adaptive procedure.

Our second result is a two-stage procedure where the first stage has controlled FDR. First introduce, for a fixed constant $\kappa \geq 2$, the following function: for $x \in [0, 1]$,

$$F_\kappa(x) = \begin{cases} 1 & \text{if } x \leq \kappa^{-1} \\ \frac{2\kappa^{-1}}{1-\sqrt{1-4(1-x)\kappa^{-1}}} & \text{otherwise} \end{cases}. \tag{7}$$

If $R_0$ denotes the first stage, we propose using $F_\kappa(|R_0|)$ as an (under-)estimation of $\pi_0^{-1}$ at the second stage. We obtain the following result:

**Theorem 4.3.** *Let $\beta$ be a fixed shape function, and $\alpha_0, \alpha_1 \in (0, 1)$ such that $\alpha_0 \leq \alpha_1$. Denote by $R_0$ the step-up procedure with threshold collection $\Delta_0(i) = \alpha_0\beta(i)/m$. Then the adaptive step-up procedure $R$ with data-dependent threshold collection $\Delta_1(i) = \alpha_1\beta(i)F_\kappa(|R_0|/m)/m$ has FDR upper bounded by $\alpha_1 + \kappa\alpha_0$ in either of the following dependence situations:*

- *the p-values $(p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$ and the shape function is the identity function.*
- *the p-values have unspecified dependencies and $\beta$ is a shape function of the form (3).*

For instance, in the PRDS case, the procedure $R$ of Theorem 4.3 with $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$, is the adaptive linear step-up procedure at level $\alpha/2$ with the following estimator for $\pi_0^{-1}$:

$$\frac{1}{1 - \sqrt{(2|R_0|/m - 1)_+}},$$

where $|R_0|$ is the number of rejections of the LSU procedure at level $\alpha/4$ and $(\cdot)_+$ denotes the positive part.

Whether in the PRDS or arbitrary dependences case, with the above choice of parameters, we note that $R$ is less conservative than the non-adaptive step-up procedure with threshold collection $\Delta(i) = \alpha\beta(i)/m$ if $F_2(|R_0| / |\mathcal{H}|) \geq 2$ or equivalently when $R_0$ rejects more than $F_2^{-1}(2) = 62,5\%$ of the null hypotheses. Conversely, $R$ is more conservative otherwise, and we can lose up to a factor 2 in the threshold collection with respect to the standard one-stage version. Therefore, here again this adaptive procedure is only useful in the cases where it is expected that a "large" proportion of null hypotheses can easily be rejected. In particular, when we use Theorem 4.3 in the distribution-free case, it is relevant to choose the shape

function $\beta$ from a prior distribution $\nu$ concentrated on the large numbers of $\{1, \dots, m\}$. Finally, note that it is not immediate to see if this procedure will improve on the one of Theorem 4.2. Namely, with the above choice parameters, it has to reject more hypotheses in the first step than the procedure of Theorem 4.2 in order to beat the LSU, and the first step is performed at a smaller target level. However, since the first step only controls the FDR, and not the FWER, it can actually be much less conservative.

To explore this issue, we performed limited experiments in a favorable situation to test the two above procedures, i.e. with a small $\pi_0$. Namely, we considered the simulation setting of Section 3.3 with $\rho = 0.1$, $m_0 = 50$ and $m = 1000$ (hence $\pi_0 = 5\%$) and $\alpha = 1\%$. The common value $\mu$ of the positive means varies in the range $[0, 5]$. Larger values of $\mu$ correspond to a very large proportion of hypotheses that are easy to reject, which favors the first stage of the two above procedures. A positively correlated family of Gaussians satisfies the PRDS assumption (see [3]), so that we use the identity shape function (linear step-up), and compare our procedures against the standard LSU. For the FWER-controlled first stage of Theorem 4.2, we chose a standard Holm procedure [18], which is a step-down procedure with threshold family $t(i) = \alpha m/(m - i + 1)$. In Figure 5, we report the average power (average number or correctly rejected false hypotheses), and the False Negative Rate (FNR), which is the converse of the FDR for type II errors, i.e., the average of the ratio of non-rejected false hypotheses over the total number of non-rejected hypotheses. Since we are in a situation where $\pi_0$ is small, the FNR might actually be a more relevant criterion than the raw power, if the goal is in fact to reliably accept a few true null hypotheses.

The conclusion is that there exists an (unfortunately quite small) region where the adaptive procedures improve over the standard LSU in terms of power. In terms of FNR, the improvement is more noticeable and over a larger region (we are not sure at this point why the FNR is a more favorable criterion for the adaptive procedures). Finallly, our two-step adaptive procedure of Theorem 4.3 appears to outperform consistently the baseline of Theorem 4.2. These results are still unsatisfying to the extent that the adaptive procedure improves over the non-adaptive one only in a region limited to some quite particular cases, and underperforms otherwise. Nevertheless, this at least demonstrates theoretically the possibility of provably adaptive procedures under dependence. Again, this theme appears to have been theoretically studied in only a handful of previous works until now, and improving significantly the theory in this setting is still an open challenge.

*Remark* 4.4. Some theoretical results for two-stage procedures under possible dependencies using a first stage with controlled FWER or controlled FDR appeared earlier in [12]. However, it seems to us that in this reference, it is implicitly assumed that the two stages are actually independent, because the proof relies on a conditioning argument wherein FDR control for the second stage still holds conditionally to the first stage output. This is the case for example is the two stages are performed on separate families of $p$-values corresponding to a new independent observation. Here we specifically wanted to take into account that we use the same collection of $p$-values for the two stages, and therefore that the two stages cannot assumed to be independent. In this sense our results are new with respect to those of [12].

*Remark* 4.5. The theoretical problem of adaptive procedures under arbitrary dependences was also addressed in [11] using two-stage procedures. However, the procedures proposed there were reported not to yield any significant improvement over non-adaptive procedures. In fact, in the explicit procedures proposed in [11], it can be seen that there exists a function $\beta$ of the form (3) such that the second stage is always more conservative than the non-adaptive

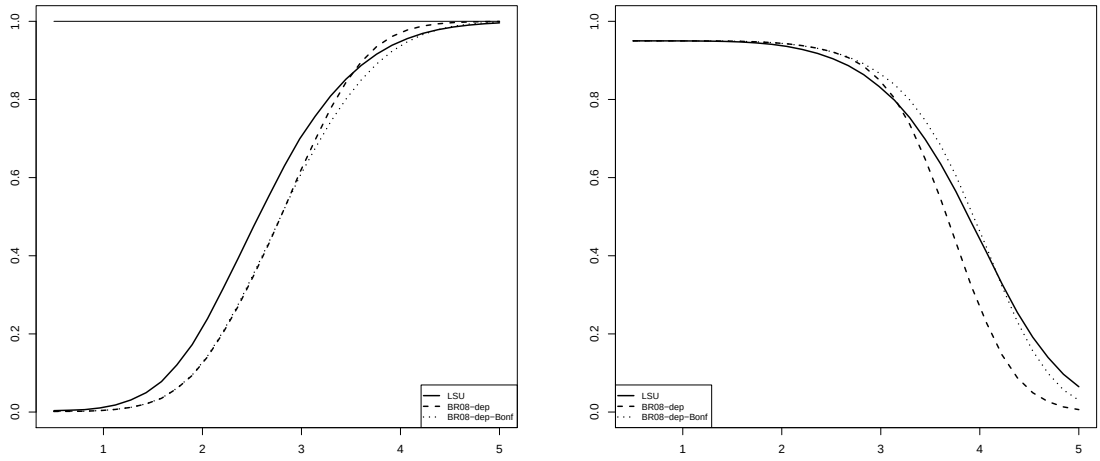


Figure 5: Average proportion of false hypotheses correctly rejected (left) and false negative rate (right) of the different procedures, as a function of the common value of alll the positive means. Parameters are $\alpha = 1\%$, $m = 1000$, $\pi_0 = 5\%$, $\rho = 0.1$. "BR08-dep-Holm" corresponds to the procedure of Theorem 4.2 using $\alpha_1 = \alpha_0 = \alpha/2$ and Holm's step-down for the first step, and "BR08-dep" to the procedure of Theorem 4.3 with $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$. The shape function $\beta$ is the identity function. Each point is an average over $10^4$ independent repetitions.

step-up procedure with threshold collection $\Delta(i) = \alpha\beta(i)/m$, (and sometimes by a large factor). Therefore, this procedure is always more conservative than the non-adaptive step-up procedure $R_\beta$ (see Theorem 2.5).

# 5 Conclusion and discussion

We proposed several adaptive multiple testing procedures that provably control the FDR under different hypotheses on the dependence of the $p$-values. First, we introduced the procedures *BR07-1S* and *BR07-2S* and we proved their theoretical validity when the $p$-values are independent. Furthermore, the procedure *BR07-2S* is less conservative in general (except in marginal situations) than the adaptive procedure proposed by [9]. Finally, extensive simulations showed that these new procedures appear to be robustly controlling the FDR even in a positive dependence situation, which is a very desirable property in practice. This is an advantage with respect to the [Storey-$\frac{1}{2}$] procedure, which is less conservative but breaks down under positive dependences. Our simulations also showed that a much smaller choice of parameter than usual in the Storey estimator, such as $\lambda = \alpha \ll 1$, resulted in a much more robust procedure under positive dependences, at the price of a slighlty more conservative procedure. This property does not appear to have been reported before. Our recommendation is therefore to replace the apparent standard default choice $\lambda = \frac{1}{2}$ in the Storey estimator by a much smaller value (typically less than 0.05).

Second, we presented what we think is among the first examples of adaptive multiple

testing procedures with provable FDR control in the PRDS and distribution-free cases. Earlier work on this topic includes [12] and [11], but an important difference is that the procedures we introduced here are both theoretically founded and can be shown to improve on non-adaptive procedures in certain circumstances. Although their interest at this point is mainly theoretical, this shows in principle that adaptivity can improve performance in a theoretically rigorous way even without the independence assumption.

Moreover, we believe that the proofs developed here are more synthetic that in the classical adaptive multiple testing literature. Indeed, we used the fact that the (adaptive) step-up procedures satisfy a form of "self-consistency condition" and then essentially just needed to apply abstract probabilistic lemmas for two dependent random variables, presented Section 7. This methodology had already allowed to simplify the proofs of classical non-adaptive results (see [16]), and avoids in particular having to reason explicitly with the reordered $p$-values, which can be somewhat cumbersome.

Another direct consequence of this approach is that the procedures investigated here in the independent case may be relatively easily extended to control the *weighted FDR*, that is, the quantity (2) where the counting measure $|.|$ has been replaced by a general measure $W(A) = \sum_{h \in A} w_h$ (with $W(\mathcal{H}) = \sum_{h \in \mathcal{H}} w_h = m$). The modifications needed to include this generalization are relatively minor. We omit the details here to avoid repetitions. This allows in particular to recover results very similar to those of [7] and can also be used to prove that a (generalized) Storey estimator can be used to control the weighted FDR, which is as far as we know a new result.

Finally, let us underline once more than the theory for adaptive procedures under dependence is still underdeveloped. It might actually be too restrictive to look for procedures having theoretically controlled FDR uniformly over very broad dependence situations such as what we studied in Section 4. An interesting future theoretical direction could be to prove that some of the adaptive procedures showing good robustness in our simulations actually have controlled FDR under some dependence, at least when the $p$-values are in some sense not too far from being independent.

# 6 Proofs of the results

## 6.1 Proofs for Section 3

**Proof of Theorem 3.1.** We denote by $R$ the corresponding procedure. Using Definition 2.3, $R$ satisfies the following relation:

$$R \subset \left\{ h \in \mathcal{H} \mid p_h \leq \min\left( (1-\lambda)\frac{\alpha|R|}{m - |R| + 1}, \lambda \right) \right\}.$$

(This type of relation was called a "self-consistency condition" in [16].) Therefore, we have

$$\begin{aligned}
\mathrm{FDR}(R) &\leq \sum_{h\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq (1-\lambda)\frac{\alpha|R(\mathbf{p})|}{m-|R(\mathbf{p})|+1}\}}{|R(\mathbf{p})|}\right] \\
&\leq \sum_{h\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq (1-\lambda)\frac{\alpha|R(\mathbf{p})|}{m-|R(\mathbf{p}_{0,h})|+1}\}}{|R(\mathbf{p})|}\right] \\
&= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq (1-\lambda)\frac{\alpha|R(\mathbf{p})|}{m-|R(\mathbf{p}_{0,h})|+1}\}}{|R(\mathbf{p})|}\Big|\mathbf{p}_{-h}\right]\right] \\
&\leq (1-\lambda)\alpha \sum_{h\in\mathcal{H}_0} \mathbb{E}\left[\frac{1}{m-|R(\mathbf{p}_{0,h})|+1}\right],
\end{aligned}$$

The last step is obtained with Lemma 7.1 of Section 7 with $U = p_h$, $g(U) = |R(\mathbf{p}_{-h}, U)|$ and $c = \frac{(1-\lambda)\alpha}{m-|R(\mathbf{p}_{0,h})|+1}$, because the distribution of $p_h$ conditionnally to $\mathbf{p}_{-h}$ is stochastically lower bounded by a uniform distribution and because $\mathbf{p}_{0,h}$ depends only on the $p$-values of $\mathbf{p}_{-h}$. Finally, since the threshold collection of $R$ is upper bounded by $\lambda$, we get

$$(1-\lambda)\mathbb{E}\left[m/(m-|R(\mathbf{p}_{0,h})|+1)\right] \leq \mathbb{E}G_1(\mathbf{p}_{0,h}),$$

where $G_1$ is the Storey estimator with parameter $\lambda$. We then use $\mathbb{E}G_1(\mathbf{p}_{0,h}) \leq \pi_0^{-1}$ (see proof of Corollary 3.5) to conclude. ■

**Proof of Lemma 3.2.** Let us denote by $\hat{t}_m$ the threshold of the LSU procedure and prove $\mathbb{P}(\hat{t}_m \leq \alpha^2 + \alpha/m) \to 0$. It was proved in [21] that under the random effects model the LSU threshold $\hat{t}_m$ converges in probability to $t^\star$, which is the largest point $t \in [0,1]$ such that $F(t) = \frac{\alpha^{-1}-\pi_0}{1-\pi_0}t$. Therefore, from the Gaussian assumption, whenever $t^* > 0$ we have

$$\mu = \overline{\Phi}^{-1}(t^\star) - \overline{\Phi}^{-1}\left(\frac{\alpha^{-1}-\pi_0}{1-\pi_0}t^\star\right).$$

It is easily seen that if $\pi_0 < (1+\alpha)^{-1}$, $\mu^*$ is well defined and for $\mu > \mu^*$, we have $t^* > \alpha^2$. Put $\varepsilon = t^\star - \alpha^2$; since for a $m$ large enough, $\mathbb{P}(\hat{t}_m \leq \alpha^2 + \alpha/m) \leq \mathbb{P}(|\hat{t}_m - t^\star| \geq \varepsilon - \alpha/m) \leq \mathbb{P}(|\hat{t}_m - t^\star| \geq \varepsilon/2)$, the first past of the result follows. Conversely, if $\pi_0 \geq (1+\alpha)^{-1}$ we have $t^* = 0$, and if $\pi_0 > (1+\alpha)^{-1}$ but $\mu < \mu^*$, we have $t^* < \alpha^2$; in both cases, by the same token as above we obtain the second part of the result. ■

**Proof of Theorem 3.4.** Denoting $R$ the procedure of Theorem 3.4 and using Definition 2.3, $R$ satisfies the following "self-consistency condition":

$$R \subset \{h \in \mathcal{H} \mid p_h \leq \alpha|R|G(\mathbf{p})/m\}. \tag{8}$$

Therefore,

$$\mathrm{FDR}(R) = \mathbb{E}\left[\frac{|R\cap\mathcal{H}_0|}{|R|}\mathbf{1}\{|R|>0\}\right] \leq \sum_{h\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha|R(\mathbf{p})|G(\mathbf{p})/m\}}{|R(\mathbf{p})|}\right].$$

Since $G$ is non-increasing, we get:

$$\begin{aligned}\mathrm{FDR}(R) &\leq \sum_{h\in\mathcal{H}_0}\mathbb{E}\left[\frac{\mathbf{1}\{p_h\leq\alpha|R(\mathbf{p})|G(\mathbf{p}_{0,h})/m\}}{|R(\mathbf{p})|}\right]\\ &= \sum_{h\in\mathcal{H}_0}\mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\{p_h\leq\alpha|R(\mathbf{p})|G(\mathbf{p}_{0,h})/m\}}{|R(\mathbf{p})|}\middle|\mathbf{p}_{-h}\right]\right]\leq\frac{\alpha}{m}\sum_{h\in\mathcal{H}_0}\mathbb{E}G(\mathbf{p}_{0,h}).\end{aligned}$$

The last step is obtained with Lemma 7.1 of Section 7 with $U=p_h$, $g(U)=|R(\mathbf{p}_{-h},U)|$ and $c=\alpha G(\mathbf{p}_{0,h})/m$, because the distribution of $p_h$ conditionnally to $\mathbf{p}_{-h}$ is stochastically lower bounded by a uniform distribution, $|R|$ is coordinate-wise non-increasing and $\mathbf{p}_{0,h}$ depends only on the $p$-values of $\mathbf{p}_{-h}$. We apply then (5) to conclude. ■

**Proof of Corollary 3.5.** By Theorem 3.4, it suffices to prove that the condition (5) holds for the nonincreasing estimators $G_i$, $i=1,2,3,4$. We reproduce here without major change the arguments used in [9] to that end. The bound for $G_1$ is obtained using Lemma 7.4 (see below) with $k=m_0$ and $q=1-\lambda$: for all $h\in\mathcal{H}_0$,

$$\mathbb{E}\left[G_1(\mathbf{p}_{0,h})\right]\leq m(1-\lambda)\mathbb{E}\left[\sum_{h'\in\mathcal{H}_0\setminus\{h\}}\mathbf{1}\{p_{h'}>\lambda\}+1\right]^{-1}\leq\pi_0^{-1}.$$

The proof for $G_2$ and $G_3$ is deduced from the one of $G_1$ because $G_2\leq G_3\leq G_1$ pointwise.

Let us prove that $\mathbb{E}G_4(\mathbf{p}_{0,h})\leq\pi_0^{-1}$, for any $h\in\mathcal{H}_0$ and any $k\in\{1,...,m\}$. If $k\leq m_1+1$, the result is trivial. Suppose now $k>m_1+1$. Introduce the following auxiliary notation: for $\mathbf{p}$ a family of $p$-values indexed by $\mathcal{H}$, and a subset $B\subset\mathcal{H}$, denote by $S(i,\mathbf{p},B)$ the $i$-th ordered $p$-value of the subfamily $(x'_h)_{h'\in B}$. Pointwise, $G_4$ can be rewritten as:

$$\begin{aligned}G_4(\mathbf{p}_{0,h}) &= \frac{m}{m+1-k}\Big(1-S(k,\mathbf{p}_{0,h},\mathcal{H})\Big)\\ &= \frac{m}{m+1-k}\Big(1-S(k-1,\mathbf{p},\mathcal{H}\setminus\{h\})\Big)\\ &\leq \frac{m}{m+1-k}\Big(1-S(k-1-m+m_0,\mathbf{p},\mathcal{H}_0\setminus\{h\})\Big),\end{aligned}$$

the latter coming from the relation $S(i,\mathbf{p},A)\geq S(i-|A\setminus B|,\mathbf{p},B)$, for every finite sets $B\subsetneq A$ and integer $i>|A\setminus B|$. Therefore, using that $m_0-1$ independent random variables with marginal distributions stochastically lower bounded by a uniform law have a $j$-largest value on average larger than $j/m_0$, we obtain:

$$\mathbb{E}G_4(\mathbf{p}_{0,h})\leq\frac{m}{m+1-k}\left(1-\frac{k-1-m+m_0}{m_0}\right)=\pi_0^{-1}.$$

■

## 6.2 Proofs for Section 4

We begin with a technical lemma that will be useful for proving both Theorem 4.2 and 4.3. It is related to techniques previously introduced in [16].

**Lemma 6.1.** *Assume $R$ is a multiple testing procedure satisfying a "self-consistency" condition of the following form:*

$$R \subset \{h \in \mathcal{H} | p_h \leq \alpha G(\mathbf{p}) \beta(|R|)/m\} \ ,$$

*where $G(\mathbf{p})$ is a data-dependent factor. Then the following inequality holds:*

$$FDR(R) \leq \alpha + \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|} \mathbf{1}\{G(\mathbf{p}) > \pi_0^{-1}\}\right] \tag{9}$$

*under either of the following conditions:*

- *the p-values $(p_h, h \in \mathcal{H})$ are PRDS on $\mathcal{H}_0$ , the cardinality $|R|$ is nonincreasing function in each p-value and and $\beta$ is the identity function.*
- *the p-values have unspecified dependencies and $\beta$ is a shape function of the form (3).*

**Proof.** We have

$$\begin{aligned}
\mathrm{FDR}(R) &= \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0}{|R|} \mathbf{1}\{|R| > 0\}\right] \\
&= \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|} \mathbf{1}\{|R| > 0\} \mathbf{1}\{G \leq \pi_0^{-1}\}\right] + \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|} \mathbf{1}\{G > \pi_0^{-1}\}\right] \\
&\leq \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq \alpha\beta(|R|)/m_0\}}{|R|}\right] + \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|} \mathbf{1}\{G > \pi_0^{-1}\}\right] .
\end{aligned}$$

The desired conclusion will therefore hold if we establish that for any $h \in \mathcal{H}_0$ , and $c > 0$ :

$$\mathbb{E}\left[\frac{\mathbf{1}\{p_h \leq c\beta(|R|)\}}{|R|}\right] \leq c \, .$$

In the distribution-free case, this is a direct consequence of Lemma 7.3 of Section 7 with $U = p_h$ and $V = \beta(|R|)$. For the PRDS case, we note that since $|R(\mathbf{p})|$ is coordinate-wise nonincreasing in each $p$-value, for any $v > 0$, $D = \{\mathbf{z} \in [0,1]^{\mathcal{H}} \mid |R(\mathbf{z})| < v\}$ is a non-decreasing set, so that the PRDS property implies that $u \mapsto \mathbb{P}(|R| < v \mid p_h = u)$ is non-decreasing. This implies that $u \mapsto \mathbb{P}(|R| < v \mid p_h \leq u)$ by the following argument (see also [22] cited in [3; 16]): putting $\gamma = \mathbb{P}\left[p_h \leq u \mid p_h \leq u'\right]$ ,

$$\begin{aligned}
\mathbb{P}\left[\mathbf{p} \in D \mid p_h \leq u'\right] &= \mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid p_h \leq u'\right] \\
&= \gamma \mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid p_h \leq u\right] + (1-\gamma)\mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid u < p_h \leq u'\right] \\
&\geq \mathbb{E}\left[\mathbb{P}\left[\mathbf{p} \in D \mid p_h\right] \mid p_h \leq u\right] = \mathbb{P}\left[\mathbf{p} \in D \mid p_h \leq u\right] .
\end{aligned}$$

We can then apply Lemma 7.2 of Section 7 with $U = p_h$ and $V = |R|$. ■

**Proof of Theorem 4.2.** By definition of a step-up procedure, the two-stage procedure $R$ satisfies the assumption of Lemma 6.1 for $G(\mathbf{p}) = (1-\frac{|R_0|}{m})^{-1}$, where $R_0$ is the first stage with FWER controlled at level $\alpha_0$. Furthermore, it is easy to check that $|R|$ is nonincreasing as a function of each $p$-value (since $|R_0|$ is). Then, we can apply Lemma 6.1, and from inequality (9) we deduce

$$\begin{aligned} FDR(R) &\leq \alpha_1 + \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{\left(1 - \frac{|R_0|}{m}\right) < \pi_0\}\right] \\ &\leq \alpha_1 + \mathbb{P}\left[R_0 \cap \mathcal{H}_0 \neq \emptyset\right] \\ &\leq \alpha_0 + \alpha_1 . \end{aligned}$$

In the case where $R_0$ rejects all hypotheses, we assumed implicitly that the second stage also does. In the above, note that if it ever happens that $R_0 = \mathcal{H}$, then either $\mathcal{H} = \mathcal{H}_0$ and the result is trivial, or this event is included in $\{R_0 \cap \mathcal{H}_0 \neq \emptyset\}$. ■

**Proof of Theorem 4.3.** Assume $\pi_0 > 0$ (otherwise the result is trivial). By definition of a step-up procedure, the two-stage procedure $R$ satisfies the assumption of Lemma 6.1 for $G(\mathbf{p}) = F_\kappa(|R_0|/m)$, where $R_0$ is the first stage. Furthermore, it is easy to check that $|R|$ is nonincreasing as a function of each $p$-value (since $|R_0|$ is). Then, we can apply Lemma 6.1, and from inequality (9) we deduce

$$\begin{aligned} \mathrm{FDR}(R) &\leq \alpha_1 + \mathbb{E}\left[\frac{|R \cap \mathcal{H}_0|}{|R|}\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}\right] \\ &\leq \alpha_1 + m_0\mathbb{E}\left[\frac{\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}}{|R_0|}\right] \end{aligned}$$

For the second inequality, we have used the two following facts:

(i) $F_\kappa(|R_0|/m) > \pi_0^{-1}$ implies $|R_0| > 0$,

(ii) because of the assumption $\alpha_0 \leq \alpha_1$ and $F_\kappa \geq 1$, the output of the second step is necessarily a set containing at least the output of the first step. Hence $|R| \geq |R_0|$.

Let us now concentrate on further bounding this second term. For this, first consider the generalized inverse of $F_\kappa$, $F_\kappa^{-1}(t) = \inf\{x \mid F_\kappa(x) > t\}$. Since $F_\kappa$ is a non-decreasing left-continuous function, we have $F_\kappa(x) > t \Leftrightarrow x > F_\kappa^{-1}(t)$. Furthermore, the expression of $F_\kappa^{-1}$ is given by: $\forall t \in [1, +\infty), F_\kappa^{-1}(t) = \kappa^{-1}t^{-2} - t^{-1} + 1$ (providing in particular that $F_\kappa^{-1}(\pi_0^{-1}) > 1 - \pi_0$). Hence

$$\begin{aligned} m_0\mathbb{E}\left[\frac{\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}}{|R_0|}\right] &\leq m_0\mathbb{E}\left[\frac{\mathbf{1}\{|R_0|/m > F_\kappa^{-1}(\pi_0^{-1})\}}{|R_0|}\right] \\ &\leq \frac{\pi_0}{F_\kappa^{-1}(\pi_0^{-1})}\mathbb{P}\left[|R_0|/m \geq F_\kappa^{-1}(\pi_0^{-1})\right] . \end{aligned} \tag{10}$$

Now, by assumption, the FDR of the first step $R_0$ is controlled at level $\pi_0\alpha_0$, so that

$$\begin{aligned} \pi_0\alpha_0 &\geq \mathbb{E}\left[\frac{|R_0 \cap \mathcal{H}_0|}{|R_0|}\mathbf{1}\{|R_0| > 0\}\right] \\ &\geq \mathbb{E}\left[\frac{|R_0| + m_0 - m}{|R_0|}\mathbf{1}\{|R_0| > 0\}\right] \\ &= \mathbb{E}\left[[1 + (\pi_0 - 1)Z^{-1}]\mathbf{1}\{Z > 0\}\right] , \end{aligned}$$

where we denoted by $Z$ the random variable $|R_0|/m$. Hence by Markov's inequality, for all $t > 1 - \pi_0$,

$$\mathbb{P}[Z \geq t] \leq \mathbb{P}\Big( [1 + (\pi_0 - 1)Z^{-1}]\mathbf{1}\{Z > 0\} \geq 1 + (\pi_0 - 1)t^{-1} \Big) \leq \frac{\pi_0 \alpha_0}{1 + (\pi_0 - 1)t^{-1}} ;$$

choosing $t = F_\kappa^{-1}(\pi_0^{-1})$ and using this into (10), we obtain

$$m_0 \mathbb{E}\left[ \frac{\mathbf{1}\{F_\kappa(|R_0|/m) > \pi_0^{-1}\}}{|R_0|} \right] \leq \alpha_0 \frac{\pi_0^2}{F_\kappa^{-1}(\pi_0^{-1}) - 1 + \pi_0} .$$

If we want this last quantity to be less than $\kappa\alpha_0$, this yields the condition $F_\kappa^{-1}(\pi_0^{-1}) \geq \kappa^{-1}\pi_0^2 - \pi_0 + 1$, and this is true from the expression of $F_\kappa^{-1}$ (note that this is how the formula for $F_\kappa$ was determined in the first place). ■

# 7 Probabilistic lemmas

The three next lemmas (Lemma 7.1, 7.2 and 7.3) have been introduced earlier in [16]. We reproduce their proof here or completeness.

**Lemma 7.1.** *Let $g : [0,1] \to (0,\infty)$ be a non-increasing function. Let $U$ be a random variable which has a distribution stochastically lower bounded by a uniform distribution, that is, $\forall u \in [0,1]$, $\mathbb{P}(U \leq u) \leq u$. Then, for any constant $c > 0$, we have*

$$\mathbb{E}\left( \frac{\mathbf{1}\{U \leq c g(U)\}}{g(U)} \right) \leq c .$$

*Proof.* We let $\mathcal{U} = \{u \mid c g(u) \geq u\}$, $u^* = \sup \mathcal{U}$ and $C^\star = \inf\{g(u) \mid u \in \mathcal{U}\}$. It is not difficult to check that $u^* \leq c C^*$ (for instance take any non-decreasing sequence $u_n \in \mathcal{U} \nearrow u^*$, so that $g(u_n) \searrow C^*$). If $C^* = 0$, then $u^* = 0$ and the result is trivial. Otherwise, we have

$$\mathbb{E}\left( \frac{\mathbf{1}\{U \leq c g(U)\}}{g(U)} \right) \leq \frac{\mathbb{P}(U \in \mathcal{U})}{C^*} \leq \frac{\mathbb{P}(U \leq u^*)}{C^*} \leq \frac{u^*}{C^*} \leq c.$$

□

**Lemma 7.2.** *Let $U, V$ be two non-negative real variables. Assume the following:*

1. *The distribution of $U$ is stochastically lower bounded by a uniform distribution, that is, $\forall u \in [0,1]$, $\mathbb{P}(U \leq u) \leq u$.*
2. *The conditional distribution of $V$ given $U \leq u$ is stochastically decreasing in $u$, that is,*

$$\forall\, v \geq 0 \qquad \forall\, 0 \leq u \leq u' , \qquad \mathbb{P}(V < v \mid U \leq u) \leq \mathbb{P}(V < v \mid U \leq u') .$$

*Then, for any constant $c > 0$, we have*

$$\mathbb{E}\left( \frac{\mathbf{1}\{U \leq c V\}}{V} \right) \leq c .$$

*Proof.* Fix some $\varepsilon > 0$ and some $\rho \in (0,1)$ and choose $K$ big enough so that $\rho^K < \varepsilon$. Put $v_0 = 0$ and $v_i = \rho^{K+1-i}$ for $1 \le i \le 2K+1$. Therefore,

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathbf{1}\{U \le cV\}}{V \vee \varepsilon}\right) &\le \sum_{i=1}^{2K+1} \frac{\mathbb{P}(U \le cv_i; V \in [v_{i-1}, v_i))}{v_{i-1} \vee \varepsilon} + \varepsilon \\
&\le c \sum_{i=1}^{2K+1} \frac{\mathbb{P}(U \le cv_i; V \in [v_{i-1}, v_i))}{\mathbb{P}(U \le cv_i)} \frac{v_i}{v_{i-1} \vee \varepsilon} + \varepsilon \\
&\le c\rho^{-1} \sum_{i=1}^{2K+1} \mathbb{P}(V \in [v_{i-1}, v_i) \mid U \le cv_i) + \varepsilon \\
&= c\rho^{-1} \sum_{i=1}^{2K+1} \big(\mathbb{P}(V < v_i \mid U \le cv_i) - \mathbb{P}(V < v_{i-1} \mid U \le cv_i)\big) + \varepsilon \\
&\le c\rho^{-1} \sum_{i=1}^{2K+1} \big(\mathbb{P}(V < v_i \mid U \le cv_i) - \mathbb{P}(V < v_{i-1} \mid U \le cv_{i-1})\big) + \varepsilon \\
&\le c\rho^{-1} + \varepsilon\,.
\end{aligned}$$

We obtain the conclusion by letting $\rho \to 1$, $\varepsilon \to 0$ and applying the monotone convergence theorem. □

**Lemma 7.3.** *Let $U, V$ be two non-negative real variables and $\beta$ be a function of the form (3). Assume that the distribution of $U$ is stochastically lower bounded by a uniform distribution, that is, $\forall u \in [0,1]$, $\mathbb{P}(U \le u) \le u$. Then, for any constant $c > 0$, we have*

$$\mathbb{E}\left(\frac{\mathbf{1}\{U \le c\beta(V)\}}{V}\right) \le c\,.$$

*Proof.* First note that since $\beta(0) = 0$, the expectation is always well defined. Since for any $z > 0$, $\int_0^{+\infty} v^{-2}\mathbf{1}\{v \ge z\}dv = 1/z$ and so using Fubini's theorem:

$$\begin{aligned}
\mathbb{E}\left(\frac{\mathbf{1}\{U \le c\beta(V)\}}{V}\right) &= \mathbb{E}\Big(\int_0^{+\infty} v^{-2}\mathbf{1}\{v \ge V\}\mathbf{1}\{U \le c\beta(V)\}dv\Big) \\
&= \int_0^{+\infty} v^{-2}\mathbb{E}[\mathbf{1}\{v \ge V\}\mathbf{1}\{U \le c\beta(V)\}]dv \\
&\le \int_0^{+\infty} v^{-2}\mathbb{P}\big(U \le c\beta(v)\big)dv \\
&\le c\int_0^{+\infty} v^{-2}\beta(v)dv,
\end{aligned}$$

and we conclude because any function $\beta$ of the form (3) satisfies $\int_0^{+\infty} v^{-2}\beta(v)dv = 1$. □

The following lemma was stated in [9]. It is a major point when we estimate $\pi_0^{-1}$ in the independent case. The proof is left to the reader.

**Lemma 7.4.** *For all $k \ge 2$, $q \in ]0,1]$ and any random variable $Y$ with a Binomial $(k-1, q)$ distribution, we have*

$$\mathbb{E}[1/(1+Y)] \le 1/kq.$$

## References


[1] Dudoit, S., Shaffer, J.P., Boldrick, J.C.: Multiple hypothesis testing in microarray experiments. Statist. Sci. **18**(1) (2003) 71–103

[2] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B **57**(1) (1995) 289–300

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

[4] Blanchard, G., Fleuret, F.: Occam's hammer. In Bshouty, N., Gentile, C., eds.: Proceedings of the 20th. conference on learning theory (COLT 2007). Volume 4539 of Springer Lecture Notes on Computer Science. (2007) 112–126

[5] Benjamini, Y., Hochberg, Y.: On the adaptive control of the false discovery fate in multiple testing with independent statistics. J. Behav. Educ. Statist. **25** (2000) 60–83

[6] Black, M.A.: A note on the adaptive control of false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. **66**(2) (2004) 297–304

[7] Benjamini, Y., Ruth, H.: False discovery rates for spatial signals. Technical report, Tel-Aviv university (2006)

[8] Storey, J.D.: A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. **64**(3) (2002) 479–498

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

[10] Storey, J.D., Taylor, J.E., Siegmund, D.: 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) (2004) 187–205

[11] Sarkar, S.K.: Two-stage stepup procedures controlling FDR. Journal of Statistical Planning and Inference **138**(4) (2008) 1072–1084

[12] Farcomeni, A.: Some results on the control of the false discovery rate under dependence. Scandinavian Journal of Statistics **34**(2) (2007) 275–297

[13] Genovese, C., Wasserman, L.: A stochastic process approach to false discovery control. Ann. Statist. **32**(3) (2004) 1035–1061

[14] Finner, H., Dickhaus, R., Roters, M.: On the false discovery rate and an asymptotically optimal rejection curve. Ann. Statist. To appear.

[15] Neuvial, P.: Asymptotic properties of false discovery rate controlling procedures under independence. ArXiv preprint math.ST/0803.2111v1 (2008)

[16] Blanchard, G., Roquain, E.: Self-consistent multiple testing procedures. ArXiV preprint math.ST/0802.1406v1 (2008)

[17] Finner, H., Roters, M.: On the false discovery rate and expected type I errors. Biom. J. **43**(8) (2001) 985–1005

[18] Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Statist. **6**(2) (1979) 65–70

[19] Storey, J.D., Tibshirani, R.: SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In: The analysis of gene expression data. Stat. Biol. Health. Springer, New York (2003) 272–290

[20] Efron, B., Tibshirani, R., Storey, J.D., Tusher, V.: Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. **96**(456) (2001) 1151–1160

[21] Genovese, C., Wasserman, L.: Operating characteristics and extensions of the false discovery rate procedure. J. R. Stat. Soc. Ser. B Stat. Methodol. **64**(3) (2002) 499–517

[22] Lehmann, E.L.: Some concepts of dependence. Ann. Math. Statist. **37** (1966) 1137–1153