Journal of Machine Learning Research 10 (2009) 2837-2871 Submitted 4/08; Revised 2/09; Published 12/09

Adaptive False Discovery Rate Control under Independence and Dependence

Gilles Blanchard GILLES.BLANCHARD@WIAS-BERLIN.DE
Weierstraß Institut für angewandte Analysis und Stochastik
Mohrenstrasse 39
10117 Berlin, Germany

Étienne Roquain ETIENNE.ROQUAIN@UPMC.FR
UPMC Univ Paris 06, UMR 7599, Laboratoire de Probabilités et Modèles Aléatoires (LPMA)
4 Place Jussieu,
75252 Paris cedex 05, France

Editor: John Shawe-Taylor

Abstract

In the context of multiple hypothesis testing, the proportion π_0 of true null hypotheses in the pool of hypotheses to test often plays a crucial role, although it is generally unknown *a priori*. A testing procedure using an implicit or explicit estimate of this quantity in order to improve its efficency is called *adaptive*. 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. In a first part, assuming independence of the p-values, we present two new procedures and give a unified review of other existing adaptive procedures that have provably controlled FDR. We report extensive simulation results comparing these procedures and testing their robustness when the independence assumption is violated. The new proposed procedures appear competitive with existing ones. The overall best, though, is reported to be Storey's estimator, albeit for a specific parameter setting that does not appear to have been considered before. In a second part, we propose adaptive versions of step-up procedures that have provably controlled FDR under positive dependence and unspecified dependence of the p-values, respectively. In the latter case, while simulations only show an improvement over non-adaptive procedures in limited situations, 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.

Keywords: multiple testing, false discovery rate, adaptive procedure, positive regression dependence, p-values

1. Introduction

The topic of multiple testing, which enjoys a long history in the statistics literature, has generated a renewed, growing attention in the recent years, spurred by an increasing number of application fields, in particular bioinformatics. For example, when processing microarray data, a common 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 the number of candidate genes is large, 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 Dudoit et al., 2003).

1.1 Adaptive Multiple Testing Procedures

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 (1995) proposed the celebrated *linear step-up* (LSU) procedure, which was proved to control the FDR under the assumption of independence between the p-values. Later, it was proved (Benjamini and Yekutieli, 2001) that the LSU procedure still controls the FDR when the p-values have positive dependence (or more precisely, a specific form of positive dependence called positive regression dependence from a subset, PRDS). Under completely unspecified dependence, 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 (Blanchard and Fleuret, 2007; Blanchard and Roquain, 2008; Sarkar, 2008a,b), by showing that there is in fact a family of step-up procedures (depending on the choice of a kind of prior distribution) that control the FDR under unspecified dependence between the p-values.

However, all of these procedures, which are built in order to control the FDR at a level α, can be shown to have actually their FDR upper bounded by $\pi_0\alpha$, where π_0 is the proportion of true null hypotheses in the initial pool. Therefore, when most of the hypotheses are false (i.e., π_0 is small), these procedures are inevitably conservative, since their FDR is in actuality much lower than the fixed target α. In this context, the challenge of *adaptive control* of the FDR (e.g., Benjamini and Hochberg, 2000; Black, 2004) is to integrate an estimation of the unknown proportion π_0 in the threshold of the previous procedures and to prove that the corresponding FDR is still rigorously controlled by α.

Of course, adaptive procedures are of practical interest if it is expected that π_0 is, or can be, significantly smaller than 1. An example of such a situation occurs when using hierarchical procedures (e.g., Benjamini and Heller, 2007) which first selects some clusters of hypotheses that are likely to contain false nulls, and then apply a multiple testing procedure on the selected hypotheses. 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 number of adaptive procedures have been proposed in the recent literature and can loosely be divided into the following categories:

- *plug-in* procedures, where some initial estimator of π_0 is directly plugged in as a multiplicative level correction to the usual procedures. In some cases (e.g., Storey's estimator, see Storey, 2002), the resulting plug-in adaptive procedure (or a variation thereof) has been proved to control the FDR at the desired level (Benjamini et al., 2006; Storey et al., 2004). A variety of other estimators of π_0 have been proposed (e.g., Meinshausen and Rice, 2006; Jin and Cai, 2007; Jin, 2008); while their asymptotic consistency (as the number of hypotheses tends to infinity) is generally established, their use in plug-in adaptive procedures has not always been studied.

- *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 π_0. Different procedures following this general approach have been proposed (Benjamini et al., 2006; Sarkar, 2008a; Farcomeni, 2007); more generally, multiple-stage procedures can be considered.

- *one-stage* procedures, which perform a single round of multiple testing (generally step-up or step-down), based on a particular (deterministic) threshold collection that renders it adaptive (Finner et al., 2009; Gavrilov et al., 2009).

In addition, some works (Genovese and Wasserman, 2004; Storey et al., 2004; Finner et al., 2009) have studied the question of adaptivity to the parameter π_0 from an *asymptotic* viewpoint. In this framework, the more specific *random effects* model is—most generally, though not always—considered, in which *p*-values are assumed independent, each hypothesis has a probability π_0 of being true, and all false null hypotheses share the same alternate distribution. The behavior of different procedures is then studied under the 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 Neuvial, 2008, for a precise study of limiting behaviors of central limit type under this model, including some of the new procedures introduced in the present paper). However, we emphasize that in the present work our focus is decidedly on the nonasymptotic side, using finite samples and arbitrary alternate hypotheses.

To complete this overview, let us also mention another interesting and different direction opened up recently, that of adaptivity to the alternate distribution. If the alternate distributions are known *a priori*, the optimal testing statistics are generally likelihood ratios between each null and each alternate, which (possibly after standardization under the form of *p*-values) can be combined using a multiple testing algorithm in order to control some measure of type I error while minimizing a measure of type II error (see, e.g., Spjøtvoll, 1972, Wasserman and Roeder, 2006, Genovese et al., 2006, Storey, 2007, Roquain and van de Wiel, 2009). In situations where the alternate is unknown, though, one can hope to estimate, implicitly or explicitly, the alternate distributions from the observed data, and consequently approximate the optimal test statistics and the associated multiple testing procedure (Sun and Cai, 2007 proposed an asymptotically consistent approach to this end). Interestingly, this point of view is also intimately linked to some traditional topics in statistical learning such as classification and of optimal novelty detection (see, e.g., Scott and Blanchard, 2009). However, in the present paper we will focus on adaptivity to the parameter π_0 only.

1.2 Overview of this Paper

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

1. We introduce a new *one-stage* step-up procedure that is more powerful than the standard LSU procedure in a large range of situations, and provably controls the FDR under independence (and in a nonasymptotic sense). This procedure is called one-stage adaptive, because the estimation of π_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 by Benjamini et al. (2006), 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 dependence, and of the family of procedures introduced by Blanchard and Fleuret, 2007, under unspecified dependence). 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 of several existing adaptive step-up procedures with provable FDR control under independence. For this, we present the theoretical FDR control as a consequence of a single general theorem, which was first established by Benjamini et al. (2006). Here, we give a short self-contained proof of this result that is of independent interest. The latter is based on some tools introduced earlier (Blanchard and Roquain, 2008; Roquain, 2007), aimed at unifying FDR control proofs. Related results and tools also appear independently in Finner et al. (2009) and Sarkar (2008b).

A third goal is to compare both the existing and our new adaptive procedures in an extensive simulation study under both independence and dependence, following the simulation model and methodology used by Benjamini et al. (2006):

- Concerning the new one- and two- stage procedures with theoretical FDR control under independence, these are generally quite competitive in comparison to existing ones. However we also report that the best procedure overall (in terms of power, among procedures that are robust enough to the dependent case) appears to be the plug-in procedure based on the well-known Storey estimator (Storey, 2002) used with the somewhat nonstandard parameter setting $\lambda = \alpha$. This outcome was in part unexpected since to the best of our knowledge, this fact had never been pointed out so far (the usual default recommended choice is $\lambda = \frac{1}{2}$ and turns out to be very unstable in dependent situations); this is therefore an important conclusion of this paper regarding practical use of these procedures.

- Concerning the new two-stage procedures with theoretical FDR control under dependence, simulations show an (admittedly limited) improvement over their non-adaptive counterparts in favorable situations which correspond to what was expected from the theoretical study (i.e., large proportion of false hypotheses). The observed improvement is unfortunately not striking enough to be able to recommend using these procedures in practice.

The paper is organized as follows: in Section 2, we introduce the mathematical framework, and we recall the existing non-adaptive results for 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 compare them theoretically and 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. A conclusion is given in Section 5. Section 6 and 7 contain proofs of the results and lemmas, respectively. 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

In this paper, we stick to the traditional statistical framework for multiple testing, which we first briefly recall below.

2.1 Multiple Testing Framework

Let $(\mathcal{X},\mathfrak{X},\mathbb{P})$ be a probability space; we aim at inferring a decision on $\mathbb{P}$ from an observation x in $\mathcal{X}$ 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, where $|.|$ is the cardinality function. The underlying probability $\mathbb{P}$ being fixed, we denote $\mathcal{H}_0=\{h\in\mathcal{H}\mid\mathbb{P}\text{ satisfies }h\}$ 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. We stress that $\mathcal{H}_0$, m_0, and π_0 are unknown and implicitly depend on the unknown $\mathbb{P}$. All the results to come are always implicitly meant to hold for any generating distribution $\mathbb{P}$.

We suppose further that there exists a set of *p-value* functions $\mathbf{p}=(p_h,h\in\mathcal{H})$, 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, each p-value is obtained from a statistic Z that has a known distribution P_0 under the corresponding null hypothesis. In this case, $p_h=\overline{\Phi}_0(Z)$ satisfies (1) in general, where $\overline{\Phi}_0(z)=P_0([z,+\infty))$. Here, we are however not concerned with how these p-values are precisely constructed and only assume that they exist and are known (this is the standard setting in multiple testing).

2.2 Multiple Testing Procedure and Errors

A *multiple testing procedure* is a function

$$R:x\in\mathcal{X}\mapsto R(x)\in\mathcal{P}(\mathcal{H}),$$

such that for any $h\in\mathcal{H}$, the function $x\mapsto\mathbf{1}\{h\in R(x)\}$ is measurable. It takes as input an observation x and returns a subset of $\mathcal{H}$, corresponding to the rejected hypotheses. As it is commonly the case, we will focus here on multiple testing procedure based on p-values, that is, we will implicitly assume that R is of the form $R(\mathbf{p})$.

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, $h\in\mathcal{H}_0\cap R$. Conversely, *a type II error* occurs for h when h is false and is not rejected by R, that is $h\in\mathcal{H}_0^c\cap R^c$. Following the Neyman-Pearson general philosophy for hypothesis testing, the primary concern is to control the quantity of type I errors of a testing procedure. For this, the most traditional way is to upper bound the "Family-wise error rate" (FWER), which is the probability that one or more true null hypotheses are rejected. However, procedures with a controlled FWER are (by definition) very "cautious" not to make even 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 hindrance in practice, since it requires to collect large enough data sets 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 (Benjamini and Hochberg, 1995): the *false discovery rate* (FDR), which is the averaged proportion of true null hypotheses in the set of all the rejected hypotheses:

Definition 1 (False discovery rate) *The* false discovery rate *of a multiple testing procedure R for a generating distribution* $\mathbb{P}$ *is given by*

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

A classical aim, then, is to build procedures R with FDR upper bounded at a given, fixed level α. Of course, if we choose $R=\emptyset$, meaning that R rejects no hypotheses, trivially $\mathrm{FDR}(R)=0\leq\alpha$. Therefore, it is desirable to build procedures R satisfying $\mathrm{FDR}(R)\leq\alpha$ while at the same time having as few type II errors as possible. As a general rule, provided that $\mathrm{FDR}(R)\leq\alpha$, we want to build procedures that reject as many false hypotheses as possible. The absolute power of a multiple testing procedure is defined as the average proportion of false hypotheses correctly rejected, $\mathbb{E}\left[\left|R\cap\mathcal{H}_0^c\right|\right]/\left|\mathcal{H}_0^c\right|$. Given two procedures R and R', a particularly simple sufficient condition for R to be more powerful than R' is when R' if $R'\subset R$ holds pointwise. We will say in this case that R is *(uniformly) less conservative* than R'.

Remark 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 notation, such as in Definition 1. Note also again that the dependence of the FDR on the unknown* $\mathbb{P}$ *is implicit.*

2.3 Self-Consistency, Step-Up Procedures, FDR Control and Adaptivity

We first define a general class of multiple testing procedures called *self-consistent procedures* (Blanchard and Roquain, 2008).

Definition 3 (Self-consistency, nonincreasing procedure) *Let* $\Delta:\{0,1,\ldots,m\}\to\mathbb{R}^+$, $\Delta(0)=0$, *be a nondecreasing function called* threshold collection*; a multiple testing procedure R is said to satisfy the self-consistency condition with respect to* Δ *if the inclusion*

$$R\subset\{h\in\mathcal{H}\mid p_h\leq\Delta(|R|)\}$$

holds almost surely. Furthermore, we say that R is nonincreasing if for all $h\in\mathcal{H}$ *the function* $p_h\mapsto|R(\mathbf{p})|$ *is nonincreasing, that is if* $|R|$ *is nonincreasing in each p-value.*

The class of self-consistent procedures includes well-known types of procedures, notably step-up and step-down. The assumption that a procedure is nonincreasing, which is required in addition to self-consistency in some of the results to come, is relatively natural as a lower p-value means we have more evidence to reject the corresponding hypothesis. We will mainly focus on the *step-up* procedure, which we define now. For this, we sort the p-values in increasing order using the notation $p_{(1)}\leq\cdots\leq p_{(m)}$ and putting $p_{(0)}=0$. This order is of course itself random since it depends on the observation.

Definition 4 (Step-up procedure) *The step-up procedure with threshold collection* Δ *is defined as*

$$R = \{h \in \mathcal{H} \mid p_h \leq p_{(k)}\}, \text{ where } k = \max\{0 \leq i \leq m \mid p_{(i)} \leq \Delta(i)\}.$$

A trivial but important property of a step-up procedure is the following.

Lemma 5 *The step-up procedure with threshold collection* Δ *is nonincreasing and self-consistent with respect to* Δ.

Therefore, a result valid for any nonincreasing self-consistent procedure w.r.t. Δ holds in particular for the corresponding step-up procedure. This will be used extensively through the paper and thus should be kept in mind by the reader.

Among all procedures that are self-consistent with respect to Δ, the step-up is uniformly less conservative than any other (Blanchard and Roquain, 2008) and is therefore of primary interest. However, to recover procedures of a more general form (including step-down for instance), the statements of this paper will be preferably expressed in terms of self-consistent procedures when it is possible.

Threshold collections are generally scaled by the target FDR level α. Once correspondingly rewritten under the normalized form $\Delta(i) = \alpha\beta(i)/m$, we will call β the *shape function* for threshold collection Δ. In the particular case where the shape function β is the identity function, the procedure is called the *linear step-up* (LSU) *procedure* (at level α).

The LSU plays a prominent role in multiple testing for 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.

Theorem 6 *Suppose that the family of p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *is independent. Then any nonincreasing self-consistent procedure with respect to threshold collection* $\Delta(i) = \alpha i/m$ *has FDR upper bounded by* $\pi_0\alpha$, *where* $\pi_0 = m_0/m$ *is the proportion of true null hypotheses. (In particular, this is the case for the linear step-up procedure.) Moreover, if the p-values associated to true null hypotheses are exactly distributed like a uniform distribution, the linear step-up procedure has FDR exactly equal to* $\pi_0\alpha$.

For the specific case of the LSU, the first part of this result was proved in the landmark paper of Benjamini and Hochberg (1995); the second part was proved by Benjamini and Yekutieli (2001) and Finner and Roters (2001). Benjamini and Yekutieli (2001) extended the first part by proving that the LSU procedure still controls the FDR in the case of *p*-values with a certain form of positive dependence called *positive regression dependence from a subset* (PRDS). We skip a formal definition for now (we will get back to this topic in Section 4). The extension of these results to self-consistent procedures (in the independent as well as PRDS case) was established by Blanchard and Roquain (2008) and Finner et al. (2009).

However, when no particular assumption is made on the dependence between the *p*-values, it can be shown that the above FDR control does not hold in general. This situation is called *unspecified* or *arbitrary* dependence. A modification of the LSU was first proposed by Benjamini and Yekutieli (2001), and proved to have a controlled FDR under arbitrary dependence. This result was extended by Blanchard and Fleuret (2007) and Blanchard and Roquain (2008) (see also a related result of Sarkar, 2008a,b). Namely, it can be shown that self-consistent procedures (not necessarily nonincreasing) based on a particular class of shape functions have controlled FDR:

Theorem 7 *Under unspecified dependence of the family of p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$*, let* β *be a shape function of the form:*

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

where ν *is some fixed a priori probability distribution on* $(0, \infty)$*. Then any self-consistent procedure with respect to threshold collection* $\Delta(i) = \alpha\beta(i)/m$ *has FDR upper bounded by* $\alpha\pi_0$*.*

To recap, in all of the above cases, the FDR is actually controlled at the level $\pi_0\alpha$ instead of the target α. Hence, a direct corollary of both of the above theorems is that the step-up procedure with shape function $\beta^*(x) = \pi_0^{-1}\beta(x)$ has FDR upper bounded by α in either of the following situations:

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

Since $\pi_0 \leq 1$, using β^* always gives rise to a less conservative procedure than using β (especially when π_0 is small). However, since π_0 is unknown, the shape function β^* is not directly accessible. We therefore call the step-up procedure using β^* the *Oracle step-up procedure* based on shape function β (in each of the above cases).

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 step-up procedures using the modified shape function $G\beta$, where G is some estimator of π_0^{-1}:

Definition 8 (Plug-in adaptive step-up procedure) *Given a level* $\alpha \in (0,1)$*, a shape function* β *and an estimator* $G : [0,1]^{\mathcal{H}} \to (0, \infty)$ *of the quantity* π_0^{-1}*, the plug-in adaptive step-up procedure of shape function* β *and using estimator G (at level* α*) is defined as*

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

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

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

1. Estimate π_0^{-1} with an estimator G.
2. Take the step-up procedure of shape function $G\beta$.

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 leads 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. Adaptive Procedures with Provable FDR Control under Independence

In this section, we introduce two new adaptive procedures that provably control the FDR under independence. The first one is one-stage and does not include an explicit estimator of π_0^{-1}, hence it is not explicitly 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 under independence using a general theoretical device introduced by Benjamini et al. (2006) which we reproduce here with a self-contained and somewhat simplified proof. Finally, we compare these different approaches; first with a theoretical study of the robustness under a very specific case of maximal dependence; second by extensive simulations, where we inspect both the performance under independence and the robustness under a wide range of positive correlations.

3.1 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 (deterministic) threshold collection.

Theorem 9 *Suppose that the p-value family* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *is independent and let* $\lambda \in (0,1)$ *be fixed. Define the adaptive threshold collection*

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

Then any nonincreasing self-consistent procedure with respect to Δ *has FDR upper bounded by* α*. In particular, this is the case of the corresponding step-up procedure, denoted by BR-1S-*λ*.*

The above result is proved in Section 6. Our proof is in part based on Lemma 1 of Benjamini et al. (2006). Note that an alternate proof of Theorem 9 is established in Sarkar (2008b) without using this lemma, while nicely connecting the FDR upper-bound to the false non-discovery rate.

3.1.1 COMPARISON TO THE LSU

Below, we will mainly focus on the choice $\lambda = \alpha$, leading to the threshold collection

$$\Delta(i) = \alpha \min\left((1-\alpha)\frac{i}{m-i+1}, 1\right). \qquad (5)$$

For $i \le (m+1)/2$, the threshold (5) 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 (Holm, 1979) for FWER control. The latter is a well-known improvement of Bonferroni's procedure (which corresponds to the fixed threshold α/m), taking into account the proportion of true nulls, and defined as the step-down procedure[1] with threshold collection $\alpha/(m-i+1)$. Here we therefore

1. The step-down procedure with threshold collection Δ rejects the hypotheses corresponding to the k smallest p-values, where $k = \max\{0 \le i \le m \mid \forall j \le i,\ p_{(j)} \le \Delta(j)\}$. It is self-consistent with respect to Δ but uniformly more conservative than the step-up procedure with the same threshold collection, compare with Definition 4.

prove that this correction is suitable as well for the linear step-up procedure, in the framework of FDR control.

If r denotes the final number of rejections of the new one-stage procedure, we can interpret the ratio $\frac{(1-\alpha)m}{m-r+1}$ between the adaptive threshold and the LSU threshold at the same point as an *a posteriori* estimate for π_0^{-1}. In the next section we propose to use this quantity in a plug-in, two-stage adaptive procedure.

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 happens only when the proportion of null hypotheses rejected by the LSU procedure is positive but less than $\alpha + 1/m$ (and even in this region the ratio of the two threshold collections is never less than $(1-\alpha)$). Roughly speaking, this situation with only few rejections can only happen if there are few false hypotheses to begin with (π_0 close to 1) or if the false hypotheses are very difficult to detect (the distribution of false p-values is close to being uniform).

In the interest of being more specific, we briefly investigate this issue in the next lemma, considering the particular *Gaussian random effects* model (which is relatively standard in the multiple testing literature, see, for example, Genovese and Wasserman, 2004) in order to give 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 π_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 $\pi_0 U[0,1] + (1-\pi_0)P_1$.

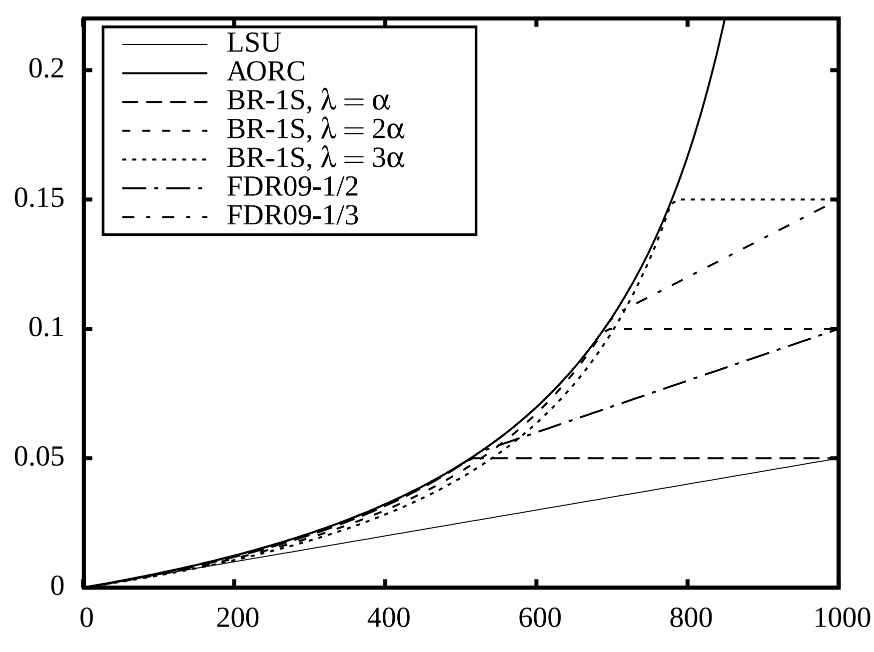

Figure 1: For $m = 1000$ null hypotheses and $\alpha = 5\%$: comparison of the new threshold collection *BR-1S*-λ given by (4) to that of the LSU, the AORC and *FDR09*-η.

Lemma 10 *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 that the true null hypotheses are standard Gaussian with zero mean and that 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.*

Lemma 10 is proved in Section 6. Taking for instance in this 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 BR-1S over the standard LSU.

3.1.2 COMPARISON TO OTHER ADAPTIVE ONE-STAGE PROCEDURES

Very recently, other adaptive one-stage procedures with important similarities to *BR-1S-*λ have been proposed by other authors. (The present work was developed independently.)

Starting with some heuristic motivations, Finner et al. (2009) proposed the threshold collection $t(i) = \frac{\alpha i}{m-(1-\alpha)i}$, which they dubbed the *asymptotically optimal rejection curve* (AORC). However, the step-up procedure using this threshold collection as is does not have controlled FDR (since $t(m) = 1$, the corresponding step-up procedure would always reject all the hypotheses), and several suitable modifications were proposed by Finner et al. (2009), the simplest one being

$$t'_\eta(i) = \min\left(t(i), \eta^{-1}\alpha i/m\right),$$

which is denoted by *FDR09-*η in the following.

The theoretical FDR control proved in Finner et al. (2009) is studied asymptotically as the number of hypotheses grows to infinity. In that framework, asymptotical control at level α is shown to hold for any $\eta < 1$. On Figure 1, we represented the thresholds *BR-1S-*λ and *FDR09-*η for comparison, for several choices of the parameters. The two families appear quite similar, initially following the AORC curve, then branching out or capping at a point depending on the parameter. One noticeable difference in the initial part of the curve is that while *FDR09-*η exactly coincides with the AORC, *BR-1S-*λ is arguably sligthly more conservative. This reflects the nature of the corresponding theoretical result—nonasymptotic control of the FDR requires a somewhat more conservative threshold as compared to the only asymptotic control of *FDR-*η. Additionally, we can use *BR-1S-*λ as a first step in a 2-step procedure, as will be argued in the next section.

The ratio between *BR-1S-*λ and the AORC (before the capping point) is a factor which, assuming $\alpha \geq (m+1)^{-1}$, is lower bounded by $(1-\lambda)(1-\frac{1}{m+1})$. This suggests that the value for λ should be kept small, this is why we propose $\lambda = \alpha$ as a default choice.

Finally, the *step-down* procedure based on the AORC threshold collection (under the slightly modified form $\widetilde{t}(i) = \frac{\alpha i}{m-(1-\alpha)i+1}$, but with no further modification) is proposed and studied by Gavrilov et al. (2009). Using specific properties of step-down procedures, these authors proved the nonasymptotic FDR control of this procedure.

3.2 Adaptive Plug-In Methods

In this section, we consider different adaptive step-up procedures of the plug-in type, that is, based on an explicit estimator of π_0^{-1}. We first review a general method proposed by Benjamini et al. (2006) in order to derive FDR control for such plug-in procedures (see also Theorem 4.3 of Finner et al., 2009, for a similar result, as well as Theorem 3.3 of Sarkar, 2008b). We propose here a self-contained proof of this result, which notably extends the original result from step-up procedures to more general self-consistent procedures. Based on this result, we review the different plug-in estimators considered by Benjamini et al. (2006) and add a new one to the lot, based on the one-stage adaptive procedure introduced in the previous section.

Let us first introduce the following notation: 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 11 (Benjamini, Krieger, Yekutieli 2006) *Suppose that the p-value family $\mathbf{p} = (p_h, h \in \mathcal{H})$ is independent. Let $G : [0,1]^{\mathcal{H}} \to (0,\infty)$ be a measurable, coordinate-wise nonincreasing function. Consider a nonincreasing multiple testing procedure R which is self-consistent with respect to the adaptive linear threshold collection $\Delta(\mathbf{p}, i) = \alpha G(\mathbf{p}) i/m$. Then the following holds:*

$$\mathrm{FDR}(R) \leq \frac{\alpha}{m} \sum_{h \in \mathcal{H}_0} \mathbb{E}\left[G(\mathbf{p}_{0,h})\right]. \tag{6}$$

In particular, if for any $h \in \mathcal{H}_0$, it holds that $\mathbb{E}\left[G(\mathbf{p}_{0,h})\right] \leq \pi_0^{-1}$, then $\mathrm{FDR}(R) \leq \alpha$.

The proof is given in Section 6. Since we assumed G to be nonincreasing, the quantity $\mathbb{E}\left[G(\mathbf{p}_{0,h})\right]$ in bound (6) is maximized when the p-values associated to true nulls have a uniform distribution (p_h excepted), while the p-values associated to false nulls are all set to zero. Following Finner et al. (2009), this *least favorable configuration* for the distribution of p-values is referred to as the Dirac-Uniform distribution and gives rise to the following corollary:

Corollary 12 *Consider the same conditions as for Theorem 11, and assume moreover that G is invariant by permutation of the p-values. Then it holds that*

$$\mathrm{FDR}(R) \leq \gamma(G, m)\, \alpha,$$

with $\gamma(G,m) = \max_{1 \leq m_0 \leq m} \left\{ \frac{m_0}{m} \mathbb{E}_{\mathbf{p} \sim DU(m, m_0 - 1)}\left[G(\mathbf{p})\right] \right\}$, where $DU(m, j)$ is the distribution of $\mathbf{p}$ where the j first p-values are independent uniform in $[0,1]$ and the $m - j$ others are identically equal to zero.

(While the proof is standard, it is given for completeness in Section 6). The interest of the last result is that for *any* choice of nonincreasing (permutation invariant) function G, it is possible in principle to evaluate $\gamma(G, m)$ by a Monte Carlo method, namely by estimating the expected value of G under the $m - 1$ possible least favorable configurations. This leads to a practical control of the FDR valid for any value of m_0, obtained by dividing the target level α by $\gamma(G, m)$ before applying the procedure.

However, when m is large, this method can be computationally demanding, and a more convenient approach for practical use is to obtain explicit bounds for specific estimators. We now concentrate on this goal and apply the result of Theorem 11 (or alternatively of Corollary 12) to the following estimators, depending on a fixed parameter $\lambda \in (0,1)$ or $k_0 \in \{1,\ldots,m\}$:

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

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

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

$$\text{[BR-2S-}\lambda\text{]} \qquad G_4(\mathbf{p}) = \frac{(1-\lambda)m}{m-|R_0'(\mathbf{p})|+1}, \text{ where } R_0' \text{ is BR-1S-}\lambda \text{ (see Theorem 9).}$$

Above, the notation "Storey-λ", "Quant-$\frac{k_0}{m}$", "BKY06-λ" and "BR-2S-λ" refer to the plug-in adaptive linear step-up procedures associated to G_1, G_2, G_3 and G_4, respectively.

Estimator G_1 is usally called *modified Storey's estimator* and was initially introduced by Storey (2002) from an heuristics on the p-values histogram (originally without the "+1", hence the name "modified"). Its intuitive justification is as follows: denoting by S_λ the set of p-values larger than the threshold λ, the average number of true nulls having a p-value in S_λ is $m_0(1-\lambda)$. Hence, a natural estimator of π_0^{-1} is $(1-\lambda)m/|S_\lambda \cap \mathcal{H}_0| \geq (1-\lambda)m/|S_\lambda| \simeq G_1(\mathbf{p})$. In particular, we expect that Storey's estimator is generally an underestimate of π_0^{-1}, which is in accordance with the condition of Theorem 11. A standard choice is $\lambda = 1/2$ (as in the SAM software of Storey and Tibshirani, 2003). FDR control for the corresponding plug-in step-up procedure was proved by Storey et al. (2004) (more precisely, for the modification $\widetilde{\Delta}(\mathbf{p},i) = \min(\alpha G_1(\mathbf{p})i/m, \lambda)$) and by Benjamini et al. (2006).

Estimator G_2 was introduced by Benjamini and Hochberg (2000) and Efron et al. (2001), from a slope heuristics on the p-values c.d.f. Roughly speaking, G_2 appears as Storey's estimator with the data-dependent parameter choice $\lambda = p_{(k_0)}$, and can therefore be interpreted as the quantile version of Storey's estimator. A standard value for k_0 is $\lfloor m/2 \rfloor$, resulting in the so-called median adaptive LSU (see Benjamini et al., 2006, and the references therein).

Estimator G_3 was introduced by Benjamini et al. (2006) for the particular choice $\lambda = \alpha/(1+\alpha)$. More precisely, a slightly less conservative version, without the "+1" in the denominator, was used in Benjamini et al. (2006). 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 π_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 $G_3 \leq G_4$ pointwise and our estimator improves over the one of Benjamini et al. (2006).

These different estimators all satisfy the sufficient condition mentioned in Theorem 11, and we thus obtain the following corollary:

Corollary 13 *Assume that the family of p-values* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *is independent. For* $i = 1,2,3,4$, *and any* $h \in \mathcal{H}_0$, *it holds that* $\mathbb{E}[G_i(\mathbf{p}_{0,h})] \leq \pi_0^{-1}$. *Therefore, the plug-in adaptive linear step-up procedure at level* α *using estimator* G_i *has FDR smaller than or equal to* α.

The above result for G_1, G_2 and G_3 (for the specific parameter setting $\lambda = \alpha/(1+\alpha)$) was proved by Benjamini et al. (2006). In Section 6, we shortly reproduce their arguments, and prove the result for G_4.

To sum up, Corollary 13 states that under independence, for any λ and k_0, the plug-in adaptive procedures Storey-λ, Quant-$\frac{k_0}{m}$, BKY06-λ and BR-2S-λ all control the FDR at level α.

Remark 14 *The result proved by Benjamini et al. (2006) is actually slightly sharper than Theorem 11. Namely, if $G(\cdot)$ is moreover supposed to be coordinate-wise left-continuous, it is possible to prove that Theorem 11 still holds when $\mathbf{p}_{0,h}$ in the RHS of (6) is replaced by the slightly better $\widetilde{\mathbf{p}}_h = (\mathbf{p}_{-h}, \widetilde{p}_h(\mathbf{p}_{-h}))$, defined as the collection of p-values $\mathbf{p}$ where p_h has been replaced by $\widetilde{p}_h(\mathbf{p}_{-h}) = \max\left\{p \in [0,1] \,\middle|\, p \le \alpha |R(\mathbf{p}_{-h}, p)| G(\mathbf{p}_{-h}, p)\right\}$. 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 15 *The one-stage step-up procedure of Finner et al. (2009) (see previous discussion in Section 3.1.2)—for which there is no result proving nonasymptotic 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 that we want to reject at least i null hypotheses whenever $p_{(i)}$ is lower than the standard LSU threshold times the estimator G_2 wherein parameter $k_0 = i$ is used. This corresponds to the inequality $p_{(i)} \le \frac{k(1-p_{(i)})}{m-i+1}$, which, solved in $p_{(i)}$, gives the threshold collection of Finner et al. (2009). Remember from Section 3.1.2 that this threshold collection must actually be modified in order to be useful, since it otherwise always leads to reject all hypotheses. The modification leading to FDR09-η consists in capping the estimated π_0^{-1} at a level η, that is, using $\min(\eta, G_2)$ instead of G_2 in the above reasoning. In fact, the proof of Finner et al. (2009) relies on a result which is essentially a reformulation of Theorem 11 for a specific form of estimator.*

Remark 16 *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 can only happen in the situation where either π_0 is close to 1 ("sparse signal") or the alternative hypotheses are difficult to detect ("weak signal"); if such a situation is anticipated, it is more appropriate to use the regular non-adaptive LSU.*

For the Storey-λ estimator, we can control precisely the probability that such an unfavorable case arises by using Hoeffding's inequality (Hoeffding, 1963): assuming the true nulls are i.i.d. uniform on $(0,1)$ and the false nulls i.i.d. of c.d.f. $F(\cdot)$, we write by definition of G_1

$$\begin{aligned}\mathbb{P}\left[G_1(\mathbf{p}) < 1\right] &= \mathbb{P}\left[\frac{1}{m}\sum_{h\in\mathcal{H}}^{m}\left(\mathbf{1}\{p_h > \lambda\} - \mathbb{P}\left[p_h > \lambda\right]\right) > (1-\pi_0)(F(\lambda)-\lambda) - m^{-1}\right] \\ &\le \exp(-2(mc^2+1)),\end{aligned}$$

where we denoted $c = (1-\pi_0)(F(\lambda)-\lambda)$, and assumed additionally $c > m^{-1}$. The behavior of the bound mainly depends on c, which can get small only if π_0 is close to 1 (sparse signal) or $F(\lambda)$ is close to λ (weak signal), illustrating the above point. In general, provided $c > 0$ does not depend on m, the probability that the Storey procedure fails to outperform the LSU vanishes exponentially as m tends to infinity.

3.3 Theoretical Robustness of the Adaptive Procedures under Maximal Dependence

For the different procedures proposed above, the theory only provides the correct FDR control under independence between the p-values. An important issue is to know how robust this control is when dependence is present (as it is often the case in practice). However, the analytic computation of the FDR under dependence is generally a difficult task, and this issue is often tackled empirically through simulations in a pre-specified model (we will do so in Section 3.4).

In this short section, we present theoretical computations of the FDR for the previously introduced adaptive step-up procedures, under the maximally dependent model where all the p-values are in fact equal, that is $p_h \equiv p_1$ for all $h \in \mathcal{H}$ (and $m_0 = m$). It corresponds to the case where we perform m times the same test, with the same p-value. Albeit relatively trivial and limited, this case leads to very simple FDR computations and provides at least some hints concerning the robustness under dependence of the different procedures studied above.

Proposition 17 *Suppose that we observe m identical p-values* $\mathbf{p} = (p_1, \ldots, p_m) = (p_1, \ldots, p_1)$ *with* $p_1 \sim U([0,1])$ *and assume* $m = m_0$. *Then, the following holds:*

$$\begin{aligned}
\mathrm{FDR}(\textit{BR-1S-}\lambda) &= \min\big(\lambda, \alpha(1-\lambda)m\big), \\
\mathrm{FDR}(\textit{FDR09-}\eta) &= \alpha\eta^{-1}, \\
\mathrm{FDR}(\textit{Storey-}\lambda) &= \min\big(\lambda, \alpha(1-\lambda)m\big) + \big(\alpha(1-\lambda)(1+m^{-1}) - \lambda\big)_+, \\
\mathrm{FDR}(\textit{Quant-}k_0/m) &= \frac{\alpha}{(1+\alpha)-(k_0-1)m^{-1}}, \\
\mathrm{FDR}(\textit{BKY06-}\lambda) &= \mathrm{FDR}(\textit{BR-2S-}\lambda) = \mathrm{FDR}(\textit{Storey-}\lambda).
\end{aligned}$$

Interestingly, the above proposition suggests specific choices of the parameters λ, η and k_0 to ensure control of the FDR at level α under maximal dependence:

- For BR-1S-λ, putting $\lambda_2 = \alpha/(\alpha + m^{-1})$, Proposition 17 gives that $\mathrm{FDR}(\textit{BR-1S-}\lambda) = \lambda$ whenever $\lambda \le \lambda_2$. This suggests to take $\lambda = \alpha$, and is thus in accordance with the default choice proposed in Section 3.1.

- For FDR09-η, no choice of $\eta < 1$ will lead to the correct FDR control under maximal dependence. However, the larger η, the smaller the FDR in this situation. Note that $\mathrm{FDR}(\textit{FDR09-}\frac{1}{2}) = 2\alpha$.

- For Storey-λ, BKY06-λ and BR-2S-λ, putting $\lambda_1 = \alpha/(1+\alpha+m^{-1})$, we have $\mathrm{FDR} = \lambda$ for $\lambda_1 \le \lambda \le \lambda_2$. This suggests to choose $\lambda = \alpha$ within these three procedures. Furthermore, note that the standard choice $\lambda = 1/2$ for Storey-λ leads to a very poor control under maximal dependence: $\mathrm{FDR}(\textit{Storey-}\frac{1}{2}) = \min(\alpha m, 1)/2$.

- For Quant-k_0/m, we see that the value of k_0 maximizing the FDR while maintaining it below α is $k_0 = \lfloor \alpha m \rfloor + 1$. Remark also that the standard choice $k_0 = \lfloor m/2 \rfloor$ leads to $\mathrm{FDR}(\textit{Quant-}k_0/m) = 2\alpha/(1+2\alpha+2m^{-1}) \simeq 2\alpha$.

Nevertheless, we would like to underline that the above computations should be interpreted with caution, as the maximal dependence case is very specific and cannot possibly give an accurate idea of the behavior of the different procedures when the correlation between the p-values are strong but not equal to 1. For instance, it is well-known that the LSU procedure has FDR far below α for strong positive correlations, but its FDR is equal to α in the above extreme model (see Finner et al., 2007, for a comprehensive study of the LSU under positive dependence). Conversely, the FDR of some adaptive procedures can be higher under moderate dependence than under maximal dependence. This behavior appears in the simulations of the next section, illustrating the complexity of the issue.

3.4 Simulation Study

How can we compare the different adaptive procedures defined above? For a fixed λ, it holds pointwise that $G_1 \geq G_4 \geq G_3$, which shows that the adaptive procedure [Storey-λ] is always less conservative than [BR-2S-λ], itself less conservative than [BKY06-λ] (except in the marginal cases where the one-stage adaptive procedure is more conservative than the standard step-up procedure, as delineated earlier for example in Lemma 10). It would therefore appear that one should always choose [Storey-λ] and disregard the other ones. However, an important point made by Benjamini et al. (2006) 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 Benjamini et al. (2006). Let $X_i = \mu_i + \varepsilon_i$, for $i, 1 \leq i \leq m$, where ε 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 ε_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 $\overline{\mu}$ for all false hypotheses, that is, for $i, 1 \leq i \leq m_0$, $\mu_i = 0$ and for $i, m_0 + 1 \leq i \leq m$, $\mu_i = \overline{\mu}$; the p-values corresponding to the null means follow exactly a uniform distribution.

Note that the case $\rho = 1$ and $m = m_0$ (i.e., $\pi_0 = 1$) corresponds to the maximally dependent case studied in Section 3.3.

We compare the following step-up multiple testing procedures: first, the one-stage step-up procedures defined in Section 3.1:

- [BR08-1S-α] The new procedure of Theorem 9, with parameter $\lambda = \alpha$,
- [FDR09-$\frac{1}{2}$] The procedure proposed in Finner et al. (2009) and described in Section 3.1.2, with $\eta = \frac{1}{2}$.

Secondly, the adaptive plug-in step-up procedures defined in Section 3.2:

- [Median LSU] The procedure [Quant-$\frac{k_0}{m}$] with the choice $\frac{k_0}{m} = \frac{1}{2}$,
- [BKY06-α] The procedure [BKY06-λ] with the parameter choice $\lambda = \alpha$,

- [BR08-2S-α] The procedure [BR08-2S-λ] with the parameter choice $\lambda = \alpha$,
- [Storey-λ] With the choices $\lambda = 1/2$ and $\lambda = \alpha$.

Finally, we used as oracle reference [LSU Oracle], the step-up procedure with the threshold collection $\Delta(i) = \alpha i/m_0$, using "oracle" prior knowledge of π_0.

The parameter choice $\lambda = \alpha$ for [Storey-λ] comes from the relationship (delineated in Section 3.1) of G_3, G_4 to G_1, and from the discussion of the maximally dependent case in Section 3.3. Note that the procedure studied by Benjamini et al. (2006) is actually [BKY06-$\alpha/(1+\alpha)$] in our notation (up to a minor modification explained in Remark 14). Thefore, the procedure [BKY06-α] used in our simulations is not srictly the same as in Benjamini et al. (2006), but it is very close.

The three most important parameters in the simulation are the correlation coefficient ρ, the proportion of true null hypotheses π_0, and the alternative mean $\overline{\mu}$ which represents the signal-to-noise ratio, 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 (π_0, $\overline{\mu}$ and ρ, respectively), the others being kept fixed. Reported are, for the different methods: the FDR, and the power relative to the reference [LSU-Oracle]. Remember the absolute power is defined as the mean proportion of false null hypotheses that are correctly rejected; for each procedure the relative power is the ratio of its absolute power to that of [LSU-Oracle]. Each point is estimated by an average of 10^5 simulations, with fixed parameters $m = 100$ and $\alpha = 5\%$.

3.4.1 UNDER INDEPENDENCE ($\rho = 0$)

Remember that under independence of the p-values, the procedure [LSU] has a FDR equal to $\alpha\pi_0$ and that the procedure [LSU Oracle] has a FDR equal to α (provided that $\alpha \leq \pi_0$). The other procedures have their FDR upper bounded by α (in an asymptotical sense only for [FDR09-$\frac{1}{2}$]).

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 independent case, the following procedures can be consistently ordered in terms of (relative) power over the range of parameters studied here:

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

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

Next, the procedures [median-LSU] and [FDR09-$\frac{1}{2}$] appear both consistently less powerful than [Storey-$\frac{1}{2}$], and [FDR09-$\frac{1}{2}$] is additionally also consistently less powerful than [Storey-α]. Their relation to the remaining procedures depends on the parameters; both [median-LSU] and [FDR09-$\frac{1}{2}$] 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 $\overline{\mu}$ is low (i.e., the alternative hypotheses are harder to distinguish).

Concerning our one-stage procedure [BR08-1S-α], we note that it appears to be indistinguishable from its two-stage counterpart [BR08-2S-α] 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 α (see, e.g., Fig. 1; a similar qualitative explanation applies to understand the behavior of [FDR09-$\frac{1}{2}$]). This is precisely to improve on this situation that we introduced the two-stage procedure, and we see that the latter does in fact improve substantially the one-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 Benjamini et al. (2006). Generally speaking, under independence we obtain a less biased estimate for π_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 λ, there will be more false null hypotheses included in the set of *p*-values larger than λ , leading to a pessimistic bias in the estimation of π_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 π_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 Benjamini et al. (2006) and we study it next.

3.4.2 UNDER POSITIVE DEPENDENCE ($\rho > 0$)

Under positive dependence, remember that it is known theoretically from Benjamini and Yekutieli (2001) that the FDR of the procedure [LSU] (resp. [LSU Oracle]) is still bounded by $\alpha\pi_0$ (resp. α), but without equality in general. However, we do not know from a theoretical point of view if the adaptive procedures have their FDR upper bounded by α. In fact, it was pointed out by Farcomeni (2007), in another work reporting simulations on adaptive procedures, that one crucial point to this respect seems to be the variability of estimate of π_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 (except the leftmost points corresponding to $\rho = 0$).

The most striking fact is that [Storey-$\frac{1}{2}$] does not control the FDR at the desired level any longer under positive dependence, and can even be off by quite a large factor. This is in accordance with the experimental findings of Benjamini et al. (2006). 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 poor FDR control in some parameter regions, most notably when π_0 is close to 1 and positive dependence is present (see, e.g., Figure 4, bottom row). In a majority of practical situations, this is an important drawback since it is difficult to rule out *a priori* that π_0 is close to 1 (i.e., there is only a small proportion of false hypotheses), or that dependence is present. Additionally, from the inspection of the behavior of the power of [median-LSU] in Figures 2 and 4, it appears that the parameter setting $\pi_0 = 0.5$ (which is the fixed value used in Figure 3) is actually noticeably the most favorable for [median-LSU] under dependence. For other values of π_0, this procedure is often clearly outperformed in terms of power, in particular by [Storey-α] and [BR-2S-α]. (At this point we have no satisfying explanation to this peculiar "peak of power" at $\pi_0 = 0.5$ observed specifically for the [median-LSU] procedure under dependence.) For all of these reasons, our conclusion is that [median-LSU] is also not robust enough in general to be reliable.

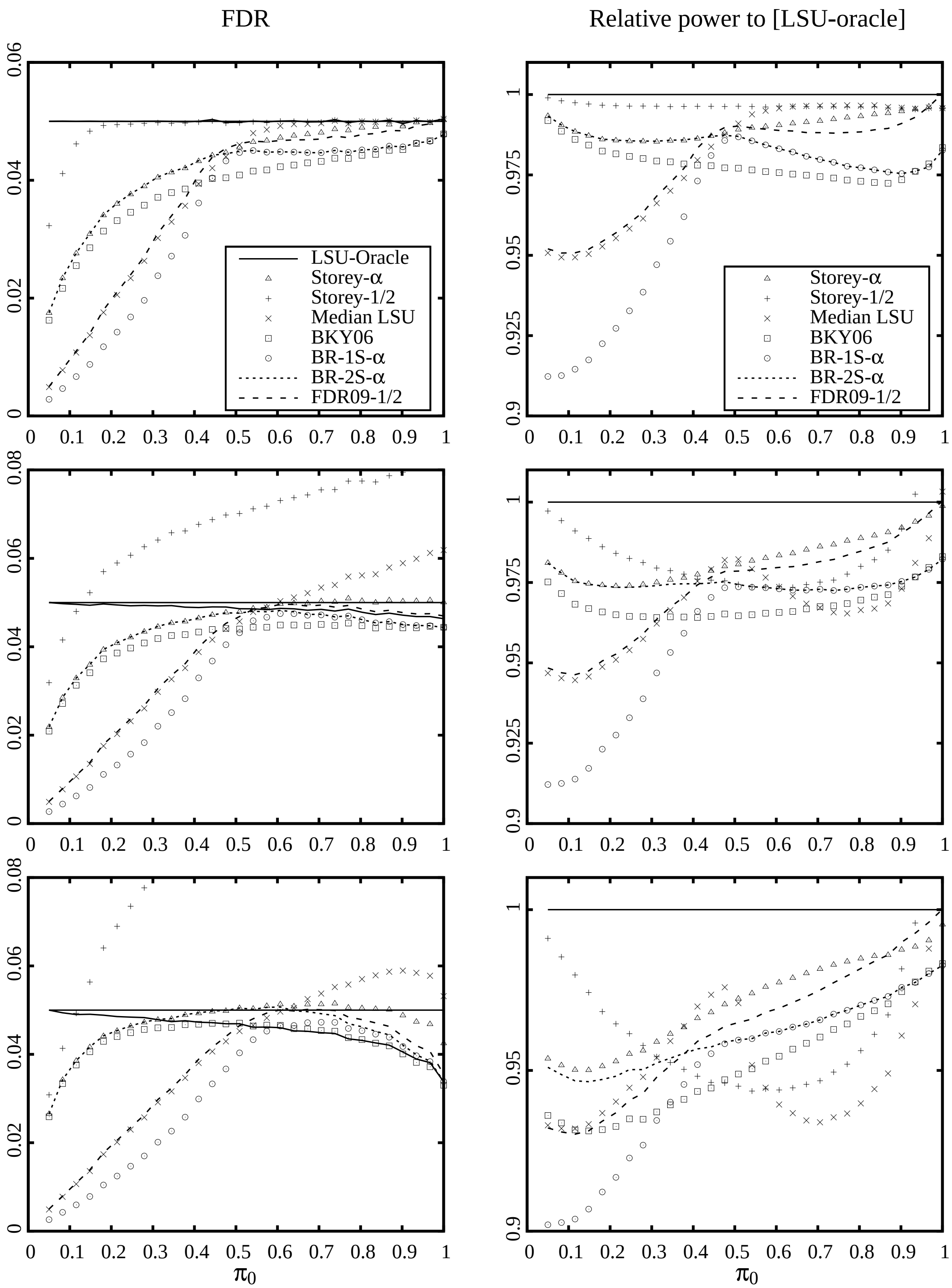

Figure 2: FDR and power relative to oracle as a function of the true proportion π_0 of null hypotheses. Target FDR is $\alpha = 5\%$, total number of hypotheses $m = 100$. The mean for the alternatives is $\overline{\mu} = 3$. From top to bottom: pairwise correlation coefficient $\rho \in \{0, 0.2, 0.5\}$.

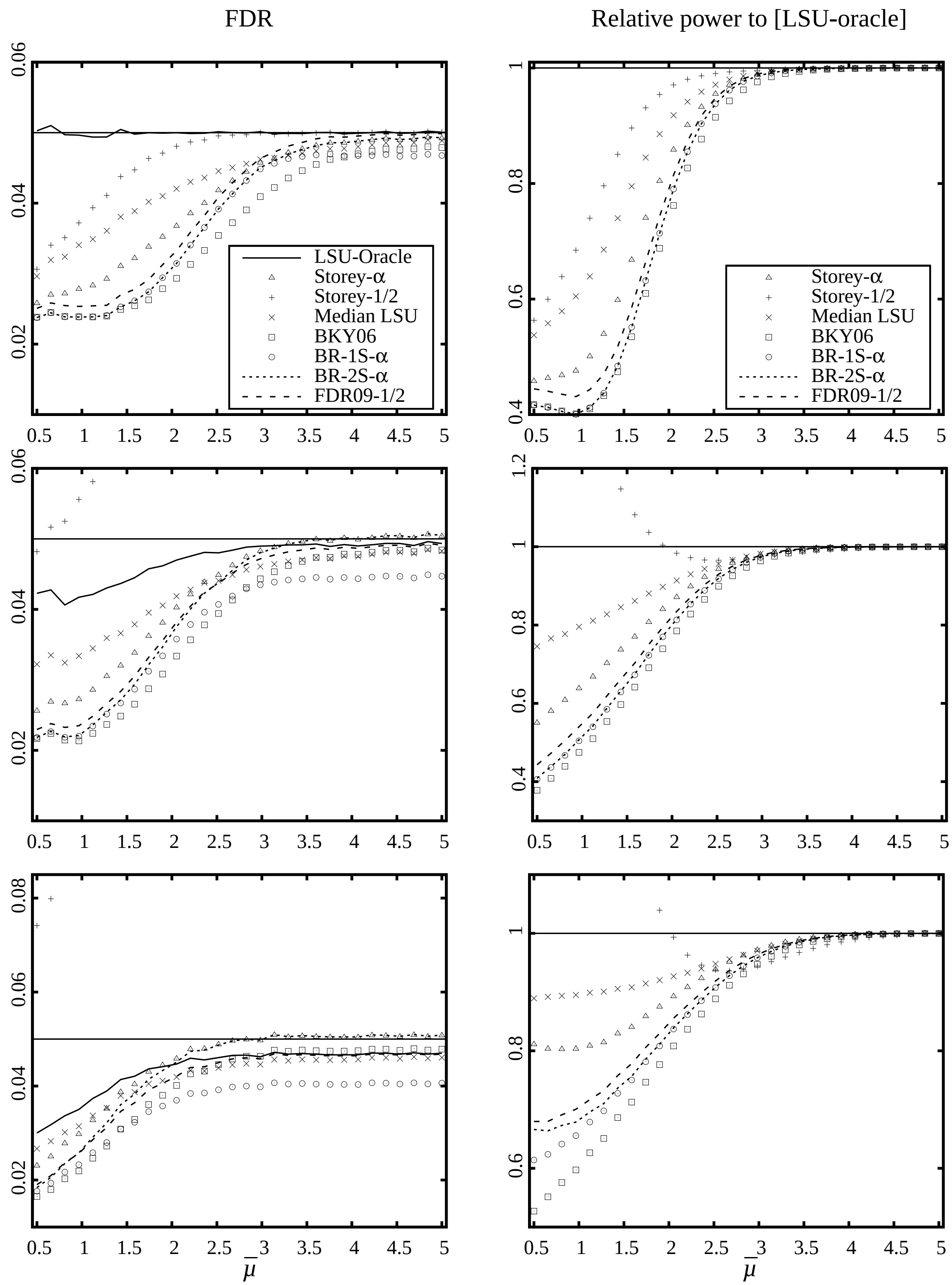

Figure 3: FDR and power relative to oracle as a function of the common alternative hypothesis mean $\bar{\mu}$. Target FDR is $\alpha = 5\%$, total number of hypotheses $m = 100$. The proportion of true null hypotheses is $\pi_0 = 0.5$. From top to bottom: pairwise correlation coefficient $\rho \in \{0, 0.2, 0.5\}$.

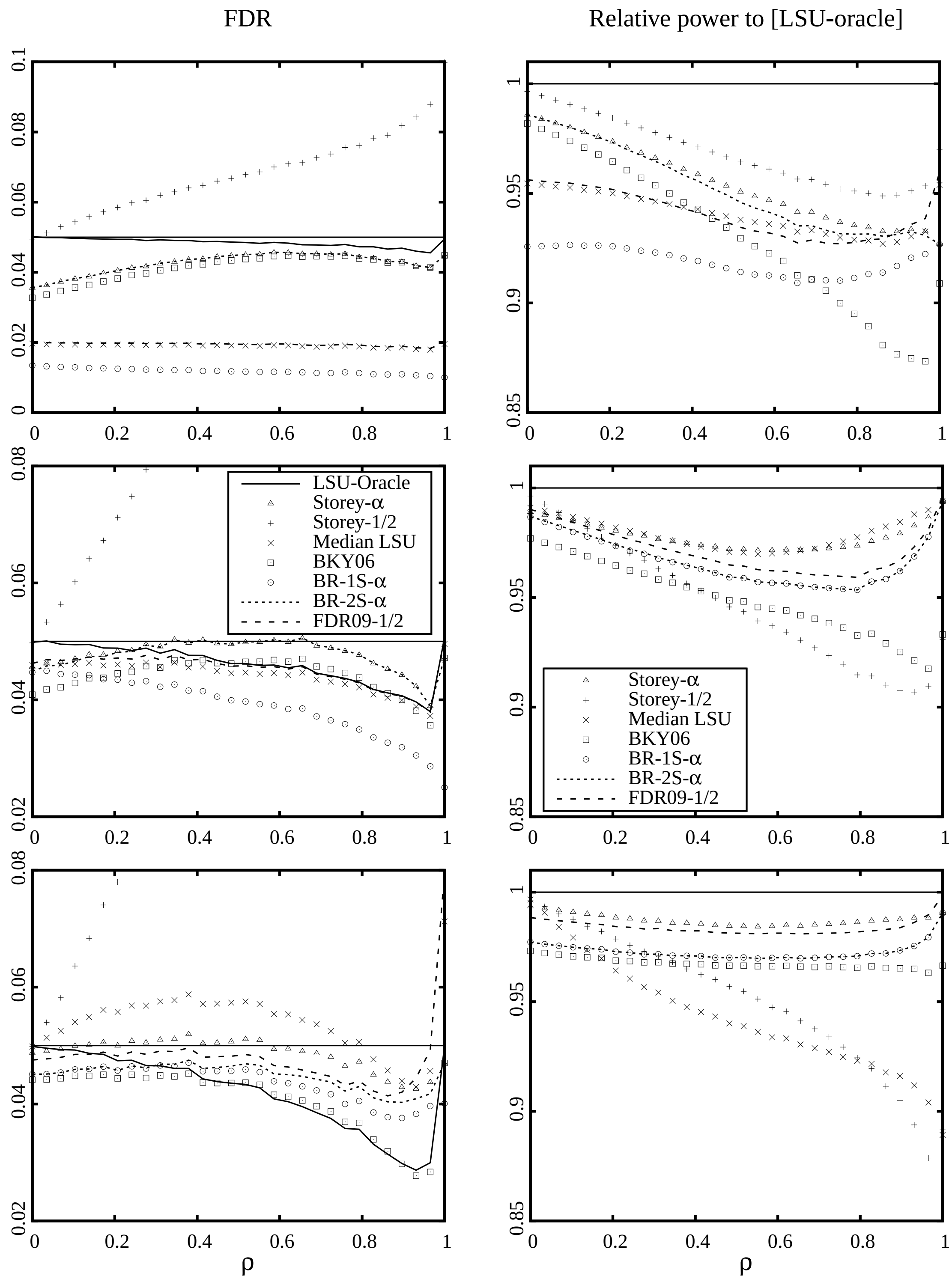

Figure 4: FDR and power relative to oracle as a function of the pairwise correlation coefficient ρ .Target FDR is $\alpha = 5\%$, total number of hypotheses $m = 100$. The mean for the alternatives is $\overline{\mu} = 3$. From top to bottom: proportion of true null hypotheses $\pi_0 \in \{0.2, 0.5, 0.8\}$.

The other remaining procedures seem to exhibit a robust control of the FDR under dependence, or at least their FDR appears to be very close to the target level (except for [FDR09-$\frac{1}{2}$] when ρ and π_0 are close to 1). For these procedures, 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-α]: 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 two-stage procedure [BR08-2S-α], then [BKY06-α].

- like in the dependent case, [FDR09-$\frac{1}{2}$] ranks second when $\pi_0 > \frac{1}{2}$ but tends to perform noticeably poorer if π_0 gets smaller. Its FDR is also not controlled if very strong correlations are present.

The overall conclusion we draw from these experiments is that for practical use, we recommend in priority [Storey-α], then as close seconds [BR08-2S-α] or [FDR09-$\frac{1}{2}$] (the latter when it is expected that $\pi_0 > 1/2$, and that there are no very strong correlations present). The procedudre [BKY06-α] is also competitive but appears to be in most cases noticeably outperformed by the above ones. These procedures all exhibit good robustness to dependence for FDR control as well as comparatively good power. The fact that [Storey-α] performs so well and seems to hold the favorite position has up to our knowledge not been reported before (it was not included in the simulations of Benjamini et al., 2006) and came somewhat as a surprise to us.

Remark 18 *As pointed out earlier, the fact that [FDR09-$\frac{1}{2}$] 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 π_0^{-1} in the procedure is capped at η (see Remark 15). Choosing a higher value for η will reduce the sub-optimality region but increase the variability of the estimate and thus decrease the overall robustness of the procedure (if dependence is present; and also under independence if only a small number m of hypotheses are tested, as for this procedure the convergence of the FDR towards its asymptotically controlled value becomes slower as η grows towards 1).*

Remark 19 *Another two-stage adaptive procedure was introduced in Sarkar (2008a), which is very similar to a plug-in procedure using [Storey-λ]. In fact, in the experiments presented in Sarkar (2008a), 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 Sarkar (2008a) 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.*

Remark 20 *One could formulate the concern that the observed FDR control for [Storey-α] could possibly fail with other parameters settings, for example when π_0 and/or ρ are close to one. We performed additional simulations to this respect (a more detailed report is available on the authors' web pages), which we summarize briefly here. We considered the following cases: $\pi_0 = 0.95$ and varying $\rho \in [0,1]$; $\rho = 0.95$ and varying $\pi_0 \in [0,1]$; finally (π_0, ρ) varying both in $[0.8,1]^2$, using a finer discretization grid to cover this region in more detail. In all the above cases Storey-α still had its FDR very close to (or below) α. Note also that the case $\rho \simeq 1$ and $\pi_0 \simeq 1$ is in accordance with* *the result of Section 3.3, stating that* $\mathrm{FDR}(Storey\text{-}\alpha) = \alpha$ *when* $\rho = 1$ *and* $\pi_0 = 1$. *Finally, we also performed additional experiments for different choices of the number of hypotheses to test (*$m = 20$ *and* $m = 10^4$*) and different choices of the target level (*$\alpha = 10\%, 1\%$*). In all of these cases were the results qualitatively in accordance with the ones already presented here.*

4. New Adaptive Procedures with Provable FDR Control under Arbitrary Dependence

In this section, we consider from a theoretical point of view the problem of constructing multiple testing procedures that are adaptive to π_0 under arbitrary dependence conditions of the p-values. The derivation of adaptive procedures that have provably controlled FDR under dependence appears to have been only studied scarcely (see Sarkar, 2008a, and Farcomeni, 2007). Here, we propose to use a two-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 π_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 the (non-adaptive) LSU procedure in the PRDS case and with respect to the family of (non-adaptive) procedures proposed in Theorem 7 in the arbitrary dependence case.

For the purposes of this section, we first recall the formal definition for PRDS dependence of Benjamini and Yekutieli (2001):

Definition 21 (PRDS condition) *Remember that a set* $D \subset [0,1]^{\mathcal{H}}$ *is said to be nondecreasing if for all* $x,y \in [0,1]^{\mathcal{H}}$*, if* $x \le y$ *coordinate-wise,* $x \in D$ *implies* $y \in D$*. Then, the p-value family* $\mathbf{p} = (p_h, h \in \mathcal{H})$ *is said to be positively regression dependent on each one from* $\mathcal{H}_0$ *(PRDS on* $\mathcal{H}_0$ *in short) if for any nondecreasing measurable set* $D \subset [0,1]^{\mathcal{H}}$ *and for all* $h \in \mathcal{H}_0$*, the function* $u \in [0,1] \mapsto \mathbb{P}[\mathbf{p} \in D \mid p_h = u]$ *is nondecreasing.*

On the one hand, it was proved by Benjamini and Yekutieli (2001) that the LSU still has controlled FDR at level $\pi_0\alpha$ (i.e., Theorem 6 still holds) under the PRDS assumption. On the other hand, under totally arbitrary dependence this result does not hold, and Theorem 7 provides a family of threshold collection resulting in controlled FDR at the same level in this case.

Our first result concerns a two-stage procedure where the first stage R_0 is any multiple testing procedure 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 22 *Let* R_0 *be a nonincreasing multiple testing procedure and assume that its FWER is controlled at level* α_0*, that is,* $\mathbb{P}[R_0 \cap \mathcal{H}_0 \neq \emptyset] \le \alpha_0$. *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-value family* $(p_h, h \in \mathcal{H})$ *is PRDS on* $\mathcal{H}_0$ *and the shape function is the identity function.*
- *the p-values have unspecified dependence and* β *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 its non-adaptive counterpart (using the same shape function) 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 $x \in [0,1]$,

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

If R_0 denotes the first stage, we propose using $F_\kappa(|R_0|/m)$ as an (under-)estimation of π_0^{-1} at the second stage. We obtain the following result:

Theorem 23 *Let β 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-value family $(p_h, h \in \mathcal{H})$ is PRDS on $\mathcal{H}_0$ and the shape function is the identity function.*
- *the p-values have unspecified dependence and β is a shape function of the form (3).*

For instance, in the PRDS case, the procedure R of Theorem 23, used with $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$, corresponds to the adaptive linear step-up procedure at level $\alpha/2$ with the following estimator for π_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 $\alpha/4$.

Whether in the PRDS or arbitrary dependence 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 $F_2(|R_0|/m) \geq 2$ 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 23 under unspecified dependence, it is relevant to choose the shape function β from a distribution ν concentrated on the large numbers of $\{1,\ldots,m\}$. Finally, note that it is not immediate to see if this procedure will improve on the one of Theorem 22. Namely, with the above choice of parameters, procedure of Theorem 22 has the advantage of using a better estimator of π_0^{-1} of the form $(1-x)^{-1} \geq (1-\sqrt{(2x-1)_+})^{-1}$ in the second round (with $x = |R_0|/m$ coming from the first round), but it has the drawback to use a first round controlling the FWER at level $\alpha/2$ which can be much more conservative than controlling the FDR at level $\alpha/4$.

To explore this issue, we performed the two above procedures, in a favorable situation where π_0 is small. Namely, we considered the simulation setting of Section 3.4 with $\rho = 0.1$, $m_0 = 100$ and $m = 1000$ (hence $\pi_0 = 10\%$) and $\alpha = 5\%$. The common value $\overline{\mu}$ of the positive means varies in the range $[0,5]$. Larger values of $\overline{\mu}$ 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 Benjamini and Yekutieli, 2001), 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 22, we chose a standard Holm procedure (see Holm, 1979), which is a step-down procedure with threshold collection $t(i) = \alpha m/(m-i+1)$. In Figure 5, we report the relative power to the oracle LSU, and the False Non-discovery Rate (FNR), which is the converse of the FDR for type II errors, that is, 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 π_0 is small, the FNR might actually be a more relevant criterion than the raw power: in this situation, because of the small number of non-rejected hypotheses, two different procedures could have their power very similar and close to 1, but noticeably different FNRs.

The conclusion is that there exists an (unfortunately relatively 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. Finally, our two-stage adaptive procedure of Theorem 23 appears to outperform consistently the baseline of Theorem 22. 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 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 24 *Some theoretical results for two-stage procedures under possible dependence using a first stage with controlled FWER or controlled FDR appeared earlier (Farcomeni, 2007). However, it appears 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 on the first stage output. This is the case for example if 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 the result of Theorem 22 is novel with respect to that of Farcomeni (2007).*

Remark 25 *The theoretical problem of adaptive procedures under arbitrary dependence was also considered by Sarkar (2008a) using two-stage procedures. However, the procedures proposed there were reported not to yield any significant improvement over non-adaptive procedures.*

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. Firstly, we introduced the one- and two-stage procedures *BR-1S* and *BR-2S* and we proved their theoretical validity when the *p*-values are independent. The procedure *BR-2S* is less conservative in general (except in marginal situations) than the adaptive procedure proposed by Benjamini et al. (2006). 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 dependence. Moreover, our simulations showed that the choice of parameter $\lambda = \alpha$ instead of $\lambda = 1/2$ in the Storey procedure

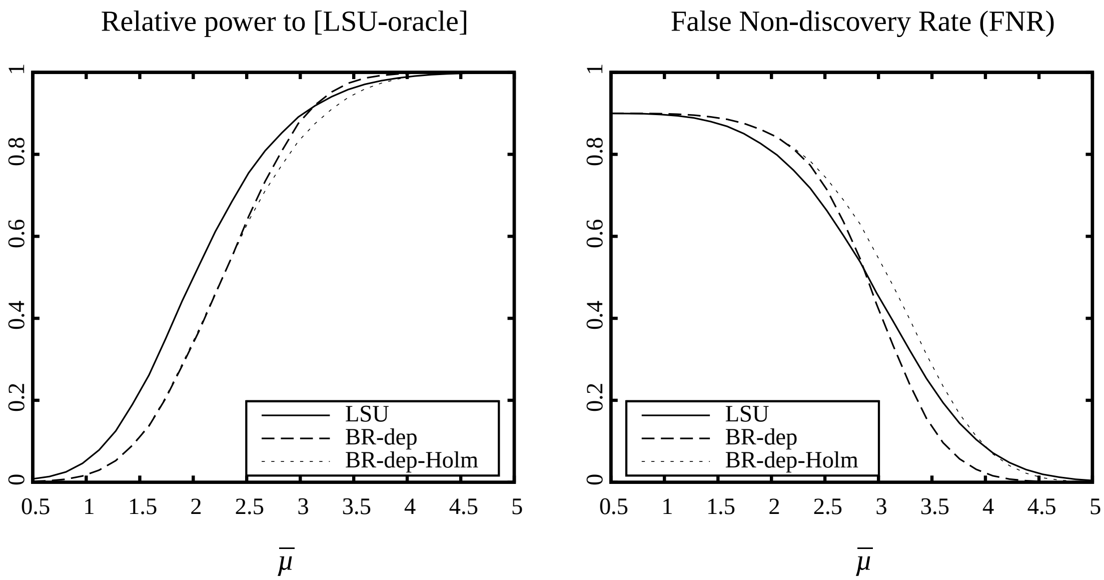

Figure 5: Relative power to oracle and false non-discovery rate (FNR) of the different procedures, as a function of the common alternative hypothesis mean $\overline{\mu}$. Parameters are $\alpha = 5\%$, $m = 1000$, $\pi_0 = 10\%$, $\rho = 0.1$. “BR08-dep-Holm” corresponds to the procedure of Theorem 22 using $\alpha_1 = \alpha_0 = \alpha/2$ and Holm’s step-down for the first step, and “BR08-dep” to the procedure of Theorem 23 with $\kappa = 2$, $\alpha_0 = \alpha/4$ and $\alpha_1 = \alpha/2$. The shape function β is the identity function. Each point is estimated by an average over 10^4 independent repetitions.

resulted in a much more robust procedure under positive dependence, at the price of being slightly more conservative. This fact is supported by a theoretical investigation of the maximally dependent case. These properties do not appear to have been reported before, and put forward Storey-α as a procedure of considerable practical interest.

Secondly, we presented what we think is among the first examples of adaptive multiple testing procedures with provable FDR control in the PRDS case and under unspecified dependence. An important difference with respect to earlier works on this topic is that the procedures we introduced here are both theoretically founded and can be shown to improve over non-adaptive procedures in certain (admittedly limited) 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.

The proofs of the results have been built upon the notion of *self-consistency* and other technical tools introduced in a previous work (Blanchard and Roquain, 2008). We believe these tools allow for a more unified approach than in the classical adaptive multiple testing literature, avoiding in particular to deal explicitly with the reordered p-values, which can be somewhat cumbersome.

Another advantage of this approach is that it can be extended in a relatively straightforward manner to the case of *weighted FDR*, that is, the quantity (2) where the cardinality measure $|.|$ has been replaced by a general measure $W(R) = \sum_{h \in R} w_h$ (with $W(\mathcal{H}) = \sum_{h \in \mathcal{H}} w_h = m$). This allows in particular to recover results very similar to those of Benjamini and Heller (2007) and can also be used to prove that a (generalized) Storey estimator can be used to control the weighted FDR. The modifications needed to include this generalizations are relatively minor; we omit the details here and refer the reader to Blanchard and Roquain (2008) to see how the case of weighted FDR can be handled using the same technical tools.

There remains a vast number of open issues concerning adaptive procedures. We first want to underline once more that 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 arbitrary 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 types of dependence, at least when the p-values are in some sense not too far from being independent.

6. Proofs

This section collects proofs for all the stated results, following their order of appearance in the text.

6.1 Proofs for Section 3

The following proofs use the notation $\mathbf{p}_{0,h}$ and $\mathbf{p}_{-h}$ defined at the beginning of Section 3.2.

6.1.1 PROOF OF THEOREM 9

Let R denote a nonincreasing self-consistent procedure with respect to Δ defined in (4). By definition, R satisfies

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

Therefore, we have

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

The second inequality above comes from $|R(\mathbf{p})| \leq |R(\mathbf{p}_{0,h})|$, which itself holds because $|R|$ is coordinate-wise nonincreasing in each p-value. The last inequality is obtained with Lemma 27 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 conditionally on $\mathbf{p}_{-h}$ is (by independence) identical to its marginal distribution, hence stochastically lower bounded by a uniform variable on $[0,1]$; $|R|$ is coordinate-wise nonincreasing; 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 λ, 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 λ. We then use $\mathbb{E}G_1(\mathbf{p}_{0,h}) \leq \pi_0^{-1}$ (see proof of Corollary 13) to conclude. ■

6.1.2 PROOF OF LEMMA 10

Denote $G(t) = \pi_0 t + (1-\pi_0)F(t)$ the c.d.f. of the p-values under the random effects mixture model. Let us denote by $\hat{t}_m$ the threshold of the LSU procedure. The proportion of rejected hypotheses from the initial pool is then exactly $\widehat{G}_m(\hat{t}_m)$, where $\widehat{G}_m$ is the empirical cdf of the p-values. It was proved by Genovese and Wasserman (2002) under the random effects model, that as m tends to infinity the LSU threshold $\hat{t}_m$ converges in probability to $t^\star$, which is the largest point $t \in [0,1]$ such that $G(t) = \alpha^{-1}t$. Since $\widehat{G}_m$ converges in probability uniformly to G, we deduce that the proportion of rejected hypotheses converges to $\alpha^{-1}t^*$ in probability; hence, if $t^* > \alpha^2$, the probability that the proportion of rejected hypotheses is less that $\alpha + 1/m$ converges to zero; and conversely converges to 1 if $t^* < \alpha^2$.

The definition of t^* and the expression for G in the Gaussian mean shift model imply the following relation whenever $t^* > 0$:

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

It is easily seen that if $\pi_0 < (1+\alpha)^{-1}$, the quantity μ^* in the statement of the lemma is well defined and we have $t^* > \alpha^2$ for $\mu > \mu^*$. This gives the first part of the result.

Conversely, if $\pi_0 > (1+\alpha)^{-1}$ we have $t^* = 0$, and if $\pi_0 < (1+\alpha)^{-1}$ but $\mu < \mu^*$, we have $t^* < \alpha^2$; this leads to the second part of the result. ■

6.1.3 PROOF OF THEOREM 11

By definition of self-consistency, the procedure R satisfies

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

Therefore,

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

Since G is nonincreasing, 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 27 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 conditionally on $\mathbf{p}_{-h}$ is (by independence) identical to its marginal distribution, hence stochastically lower bounded by a uniform variable; $|R|$ is coordinate-wise nonincreasing; and $\mathbf{p}_{0,h}$ depends only on the p-values of $\mathbf{p}_{-h}$. ■

6.1.4 PROOF OF COROLLARY 12

Assuming $\mathcal{H}_0 \neq \emptyset$ without loss of generality, for $h_0 \in \mathcal{H}_0$ we want to upper bound $\mathbb{E}[G(\mathbf{p}_{0,h_0})]$ appearing in the bound of Theorem 11. Let $\widetilde{\mathbf{p}}_{0,h_0}$ denote the family of p-values $\mathbf{p}_{0,h_0}$ where all p-values p_h, $h \in \mathcal{H} \setminus \mathcal{H}_0$ have been replaced by zero. Since G is nonincreasing, we have $\mathbb{E}[G(\mathbf{p}_{0,h_0})] \leq \mathbb{E}[G(\widetilde{\mathbf{p}}_{0,h_0})]$. Now, for any $h \in \mathcal{H}_0 \setminus \{h_0\}$, denote $\widetilde{\mathbf{p}}'_{0,h_0}$ the family $\widetilde{\mathbf{p}}_{0,h_0}$, where the variable p_h has been replaced by u_h, an independent uniform variable on $[0,1]$. Since both p_h and u_h are independent of the other p-values, p_h is stochastically lower bounded by u_h and G is nonincreasing, we have

$$\mathbb{E}\left[G(\widetilde{\mathbf{p}}_{0,h_0}) \middle| \mathbf{p}_{-h}\right] \leq \mathbb{E}\left[G(\widetilde{\mathbf{p}}'_{0,h_0}) \middle| \mathbf{p}_{-h}\right] ,$$

hence also in unconditional expectation. Iterating this reasoning in succession for all $h \in \mathcal{H}_0 \setminus \{h_0\}$, we have finally replaced $\mathbf{p}_{0,h_0}$ by a family of $m_0 - 1$ independent uniform variables and $m - m_0 + 1$ zeros, while only increasing the expected value, so that (now using that G is permutation invariant)

$$\mathbb{E}[G(\mathbf{p}_{0,h_0})] \leq \mathbb{E}_{\mathbf{p}\sim DU(m,m_0-1)}[G(\mathbf{p})] ,$$

which, combined with (6), entails the desired result. ■

6.1.5 PROOF OF COROLLARY 13

First, we prove that the sufficient condition of Theorem 11 holds for the nonincreasing estimators G_i, $i = 1,3,4$. To that end, we reproduce here without major changes the arguments used by Benjamini et al. (2006). The bound for G_1 is obtained using Lemma 30 (see below) with $k = m_0$ and $q = 1-\lambda$: for all $h \in \mathcal{H}_0$,

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

The proof for G_3 and G_4 is deduced from the one of G_1 because G_3 and G_4 are smaller than G_1 pointwise.

Secondly, for G_2 we use a somewhat more direct argument than Benjamini et al. (2006), namely using Corollary 12 and proving that $\gamma(G_2, m) \leq 1$. Take $\mathbf{p} \sim DU(m, m_0 - 1)$. On the one hand, if $k_0 \leq m - m_0 + 1$, we have $p_{(k_0)} = 0$, and therefore $\pi_0 G_2(\mathbf{p}) = \pi_0 m/(m - k_0 + 1) \leq 1$ pointwise. On the other hand, if $k_0 \geq m - m_0 + 2$, we have $p_{(k_0)} = q_{(k_0 - m + m_0 - 1)}$, where $q_{(1)} \leq ... \leq q_{(m_0-1)}$ are the $(m_0 - 1)$ ordered p-values of $\mathbf{p}$ corresponding to uniform variables. Thus,

$$\pi_0 \mathbb{E}[G_2(\mathbf{p})] = m\pi_0 \frac{1 - \mathbb{E}\left[q_{(k_0-m+m_0-1)}\right]}{m - k_0 + 1} = m\pi_0 \frac{1 - (k_0 - m + m_0 - 1)/m_0}{m - k_0 + 1} = 1.$$

6.1.6 PROOF OF PROPOSITION 17

Let us first consider adaptive one-stage procedures: for any step-up procedure R of threshold $\Delta(i) = \alpha\beta(i)/m$ we easily derive that the probability that R makes any rejection is

$$\mathbb{P}[\exists i \mid p_i \leq \Delta(i)] = \mathbb{P}[\exists i \mid p_1 \leq \Delta(i)] = \mathbb{P}[p_1 \leq \Delta(m)] = \Delta(m),$$

which is $\mathrm{FDR}(R)$ because $m_0 = m$. The results for *BR-1S*-λ and *FDR09*-η follow.

With the same reasoning, we find that for any plug-in adaptive linear step-up procedure R that uses an estimator $G(\mathbf{p})$,

$$\mathrm{FDR}(R) = \mathbb{P}\left[p_1 \le \alpha G(\mathbf{p})\right]. \tag{7}$$

Next, for the Storey plug-in procedure, we have $G_1(p_1,\ldots,p_1) = (1-\lambda)m/(m\mathbf{1}\{p_1>\lambda\}+1)$, so that applying (7), we get

$$\begin{aligned}\mathrm{FDR}(\textit{Storey-}\lambda) &= \mathbb{P}\left[p_1 \le \alpha G_1(\mathbf{p})\right]\\ &= \mathbb{P}\left[p_1 \le \lambda,\, p_1 \le \alpha(1-\lambda)m\right] + \mathbb{P}\left[p_1 > \lambda,\, p_1 \le \alpha(1-\lambda)m/(m+1)\right]\\ &= \min\Big(\lambda, \alpha(1-\lambda)m\Big) + \left(\frac{\alpha(1-\lambda)m}{m+1} - \lambda\right)_+ .\end{aligned}$$

For the quantile procedure, we have

$$\mathbb{P}\left[p_1 \le \alpha(1-p_1)m/(m-k_0+1)\right] = \mathbb{P}\left[p_1((1+\alpha)m-k_0+1) \le \alpha m\right] = \frac{\alpha}{1+\alpha-(k_0-1)/m}.$$

For the BKY06 procedure, we simply remark that since the linear step-up procedure of level λ rejects all the hypotheses when $p_1 \le \lambda$ and rejects no hypothesis otherwise, the estimator G_1 and G_3 are equal in this case. The proof for *BR-2S*-λ is similar. ■

6.2 Proofs for Section 4

We begin with a technical lemma that will be useful for proving both Theorem 22 and 23. It is related to techniques previously introduced by Blanchard and Roquain (2008).

Lemma 26 *Assume R is a multiple testing procedure satisfying the self-consistency condition:*

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

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

$$FDR(R) \le \alpha + \mathbb{E}\left[\frac{|R\cap\mathcal{H}_0|}{|R|}\mathbf{1}\{|R|>0\}\mathbf{1}\left\{G(\mathbf{p}) > \pi_0^{-1}\right\}\right], \tag{8}$$

under either of the following conditions:

- *the p-value family $(p_h, h\in\mathcal{H})$ is PRDS on $\mathcal{H}_0$, R is nonincreasing and β is the identity function.*
- *the p-values have unspecified dependence and β is a shape function of the form (3).*

Proof. We have

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

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

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

Under unspecified dependence, this is a direct consequence of Lemma 29 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 measurable nondecreasing set, so that the PRDS property implies that $u \mapsto \mathbb{P}[|R| < v \mid p_h = u]$ is nondecreasing. This implies that $u \mapsto \mathbb{P}[|R| < v \mid p_h \le u]$ by the following argument (see also Lehmann, 1966, cited by Benjamini and Yekutieli, 2001, and Blanchard and Roquain, 2008): putting $\gamma = \mathbb{P}[p_h \le u \mid p_h \le u']$,

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

We can then apply Lemma 28 of Section 7 with $U = p_h$ and $V = |R|$. ■

6.2.1 PROOF OF THEOREM 22

By definition of a step-up procedure, the two-stage procedure R satisfies the assumption of Lemma 26 for $G(\mathbf{p}) = (1 - \frac{|R_0|}{m})^{-1}$, where R_0 is the first stage with FWER controlled at level α_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 26, and from inequality (8) we deduce

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

In the case where R_0 rejects all hypotheses, we assumed implicitly that the second stage also does. ■

6.2.2 PROOF OF THEOREM 23

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 26 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 26, and from inequality (8) we deduce

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

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 $\alpha_0 \le \alpha_1$ and $F_\kappa \ge 1$, the output of the second step is necessarily a set containing at least the output of the first step. Hence $|R| \ge |R_0|$.

Let us now concentrate on further bounding this second term. For this, first consider the generalized inverse of F_κ, $F_\kappa^{-1}(t) = \inf\{x \mid F_\kappa(x) > t\}$. Since F_κ is a nondecreasing left-continuous function, we have $F_\kappa(x) > t \Leftrightarrow x > F_\kappa^{-1}(t)$. Furthermore, the expression of F_κ^{-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

$$m_0 \mathbb{E}\left[\frac{\mathbf{1}\left\{F_\kappa(|R_0|/m) > \pi_0^{-1}\right\}}{|R_0|}\right] \le m_0 \mathbb{E}\left[\frac{\mathbf{1}\left\{|R_0|/m > F_\kappa^{-1}(\pi_0^{-1})\right\}}{|R_0|}\right]$$
$$\le \frac{\pi_0}{F_\kappa^{-1}(\pi_0^{-1})} \mathbb{P}\left[|R_0|/m \ge F_\kappa^{-1}(\pi_0^{-1})\right]. \qquad (9)$$

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

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

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

$$\mathbb{P}[Z \ge t] \le \mathbb{P}\left[\left(1 + (\pi_0 - 1)Z^{-1}\right)\mathbf{1}\{Z > 0\} \ge 1 + (\pi_0 - 1)t^{-1}\right] \le \frac{\pi_0\alpha_0}{1 + (\pi_0 - 1)t^{-1}};$$

choosing $t = F_\kappa^{-1}(\pi_0^{-1})$ and using this into (9), we obtain

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

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

7. Probabilistic Lemmas

The three following lemmas have been established in a previous work (see Blanchard and Roquain, 2008, Lemma 3.2).

Lemma 27 *Let $g : [0,1] \to (0,\infty)$ be a nonincreasing function. Let U be a random variable which is stochastically lower bounded by a uniform variable on $[0,1]$, that is, $\forall u \in [0,1]$, $\mathbb{P}[U \le u] \le u$. Then, for any constant $c > 0$, we have*

$$\mathbb{E}\left[\frac{\mathbf{1}\{U \le cg(U)\}}{g(U)}\right] \le c.$$

Lemma 28 *Let U, V be two nonnegative real variables. Assume the following:*

1. *U is stochastically lower bounded by a uniform variable on $[0,1]$, 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,*

$$\forall v \geq 0, \qquad \forall 0 \leq u \leq u', \qquad \mathbb{P}[V < v \mid U \leq u] \leq \mathbb{P}\left[V < v \mid U \leq u'\right].$$

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

$$\mathbb{E}\left[\frac{\mathbf{1}\{U \leq cV\}}{V}\right] \leq c.$$

Lemma 29 *Let U, V be two nonnegative real variables and β be a function of the form (3). Assume that U is stochastically lower bounded by a uniform variable on $[0,1]$, that is, $\forall u \in [0,1]$, $\mathbb{P}[U \leq u] \leq u$. Then, for any constant $c > 0$, we have*

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

The following lemma was stated by Benjamini et al. (2006). It is a major point when we estimate π_0^{-1} in the independent case. The proof is left to the reader.

Lemma 30 *For any $k \geq 2$, $q \in (0,1]$, let Y be a binomial random variable with parameters $(k-1, q)$; then the following holds:*

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

Acknowledgments

The first author's work was carried out at the Fraunhofer Institute FIRST, Berlin. The second author's research was mostly carried out at the French institute INRA and at the VU University in Amsterdam, that are warmly acknowledged. Both authors would like to thank the Associated Editor and the two referees for their helpful comments and suggestions.

References

Y. Benjamini and R. Heller. False discovery rates for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281, 2007.

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

Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25:60–83, 2000.

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

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.

M. A. Black. A note on the adaptive control of false discovery rates. *Journal of the Royal Statistical Society, Series B*, 66(2):297–304, 2004.

G. Blanchard and F. Fleuret. Occam's hammer. In N.H. Bshouty and C. Gentile, editors, *Proceedings of the 20th. Conference on Learning Theory (COLT 2007)*, volume 4539 of *Springer Lecture Notes on Computer Science*, pages 112–126, 2007.

G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic Journal of Statistics*, 2:963–992, 2008.

S. Dudoit, J. Shaffer, and J. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

B. Efron, R. Tibshirani, J. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160, 2001.

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

H. Finner and M. Roters. On the false discovery rate and expected type I errors. *Biometrical Journal*, 43(8):985–1005, 2001.

H. Finner, T. Dickhaus, and M. Roters. Dependency and false discovery rate: Asymptotics. *Annals of Statistics*, 35(4):1432–1455, 2007.

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

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

C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society, Series B*, 64(3):499–517, 2002.

C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, 32(3):1035–1061, 2004.

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

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

J. Jin. Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators. *Journal of the Royal Statistical Society, Series B*, 70(3):461–493, 2008.

J. Jin and T. Cai. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506, 2007.

E. L. Lehmann. Some concepts of dependence. *Annals of Mathematical Statistics*, 37:1137–1153, 1966.

N. Meinshausen and J. Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Annals of Statistics*, 34(1):373–393, 2006.

P. Neuvial. Asymptotic properties of false discovery rate controlling procedures under independence. *Electronic Journal of Statistics*, 2:1065–1110, 2008.

E. Roquain. *Exceptional Motifs in Heterogeneous Sequences. Contributions to Theory and Methodology of Multiple Testing*. PhD thesis, Université Paris XI, 2007.

E. Roquain and M. van de Wiel. Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3:678–711, 2009.

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

S. K. Sarkar. On methods controlling the false discovery rate. *Sankhya, Series A*, 70:135–168, 2008b.

C. Scott and G. Blanchard. Novelty detection: unlabaled data definitely help. In D. van Dyk and M. Welling, editors, *Proceedings of the 12th. International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, volume 5 of *JMLR Workshop and Conference Proceedings*, pages 464–471, 2009.

E. Spjøtvoll. On the optimality of some multiple comparison procedures. *Annals of Mathematical Statistics*, 43(2):398–411, 1972.

J. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64(3):479–498, 2002.

J. Storey. The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society, Series B*, 69(3):347–368, 2007.

J. Storey and R. Tibshirani. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data*, Statistics for Biology and Health, pages 272–290. Springer, New York, 2003.

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

W. Sun and T. Cai. Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912, 2007.

L. Wasserman and K. Roeder. Weighted hypothesis testing. Technical report 830, Dept. of statistics, Carnegie Mellon University, 2006. ArXiv preprint math/0604172v1. URL: `http://arxiv.org/abs/math.ST/0604172`.