Online Rules for Control of False Discovery Rate and False Discovery Exceedance

Adel Javanmard[*] and Andrea Montanari[†]

August 15, 2016

Abstract

Multiple hypothesis testing is a core problem in statistical inference and arises in almost every scientific field. Given a set of null hypotheses $\mathcal{H}(n) = (H_1, \dots, H_n)$, Benjamini and Hochberg [BH95] introduced the false discovery rate (FDR), which is the expected proportion of false positives among rejected null hypotheses, and proposed a testing procedure that controls FDR below a pre-assigned significance level. Nowadays FDR is the criterion of choice for large-scale multiple hypothesis testing.

In this paper we consider the problem of controlling FDR in an *online manner*. Concretely, we consider an ordered –possibly infinite– sequence of null hypotheses $\mathcal{H} = (H_1, H_2, H_3, \dots)$ where, at each step i, the statistician must decide whether to reject hypothesis H_i having access only to the previous decisions. This model was introduced by Foster and Stine [FS07].

We study a class of *generalized alpha-investing* procedures, first introduced by Aharoni and Rosset [AR14]. We prove that any rule in this class controls online FDR, provided p-values corresponding to true nulls are independent from the other p-values. Earlier work only established mFDR control. Next, we obtain conditions under which generalized alpha-investing controls FDR in the presence of general p-values dependencies. Finally, we develop a modified set of procedures that also allow to control the false discovery exceedance (the tail of the proportion of false discoveries).

Numerical simulations and analytical results indicate that online procedures do not incur a large loss in statistical power with respect to offline approaches, such as Benjamini-Hochberg.

1 Introduction

The common practice in claiming a scientific discovery is to support such claim with a p-value as a measure of statistical significance. Hypotheses with p-values below a significance level α, typically 0.05, are considered to be *statistically significant*. While this ritual controls type I errors for single testing problems, in case of testing multiple hypotheses it leads to a large number of false positives (false discoveries). Consider, for instance, a setting in which N hypotheses are to be tested, but only a few of them, say s, are non-null. If we test all of the hypotheses at a fixed significance level α, each of $N - s$ truly null hypotheses can be falsely rejected with probability α. Therefore, the number of false discoveries –equal to $\alpha(N - s)$ in expectation– can substantially exceed the number s of true non-nulls.

[*] Department of Data Sciences and Operations, University of Southern California. Email: `ajavanma@usc.edu`

[†] Department of Electrical Engineering and Department of Statistics, Stanford University. Email: `montanar@stanford.edu`

The false discovery rate (FDR) –namely, the expected fraction of discoveries that are false positives– is the criterion of choice for statistical inference in large scale hypothesis testing problem. In their groundbreaking work [BH95], Benjamini and Hochberg (BH) developed a procedure to control FDR below a pre-assigned level, while allowing for a large number of true discoveries when many non-nulls are present. The BH procedure remains –with some improvements– the state-of-the-art in the context of multiple hypothesis testing, and has been implemented across genomics [RYB03], brain imaging [GLN02], marketing [PWJ15], and many other applied domains.

Standard FDR control techniques, such as the BH procedure [BH95], require aggregating p-values for all the tests and processing them jointly. This is impossible in a number of applications which are best modeled as an online hypothesis testing problem [FS07] (a more formal definition will be provided below):

> *Hypotheses arrive sequentially in a stream. At each step, the analyst must decide whether to reject the current null hypothesis without having access to the number of hypotheses (potentially infinite) or the future p-values, but solely based on the previous decisions.*

This is the case –for instance– with publicly available datasets, where new hypotheses are tested in an on-going fashion by different researchers [AR14]. Similar constraints arise in marketing research, where multiple A-B tests are carried out on an ongoing fashion [PWJ15]. Finally, scientific research as a whole suffers from the same problem: a stream of hypotheses are tested on an ongoing basis using a fixed significance level, thus leading to large numbers of false positives [Ioa05b]. We refer to Section 8 for further discussion.

In order to illustrate the online scenario, consider an approach that would control the family-wise error rate (FWER), i.e. the probability of rejecting at least one true null hypothesis. Formally

$$\mathrm{FWER}(n) \equiv \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{P}_{\theta}\Big(V^{\theta}(n) \geq 1\Big)\,. \tag{1}$$

where $\boldsymbol{\theta}$ denotes the model parameters (including the set of non-null hypotheses) and $V^{\theta}(n)$ the number of false positives among the first n hypotheses. This metric can be controlled by choosing different significance levels α_i for tests H_i, with $\boldsymbol{\alpha} = (\alpha_i)_{i\geq 1}$ summable, e.g., $\alpha_i = \alpha 2^{-i}$. Notice that the analyst only needs to know the number of tests performed before the current one, in order to implement this scheme. However, this method leads to small statistical power. In particular, making a discovery at later steps becomes very unlikely.

In contrast, the BH procedure assumes that all the p-values are given a priori. Given p-values $p_1, p_2, \dots, p_N$ and a significance level α, BH follows the steps below:

1. Let $p_{(i)}$ be the i-th p-value in the (increasing) sorted order, and define $p_{(0)} = 0$. Further. let

$$i_{\mathrm{BH}} \equiv \max\Big\{0 \leq i \leq N : \, p_{(i)} \leq \alpha i/N\Big\}\,. \tag{2}$$

2. Reject H_j for every test with $p_j \leq p_{(i_{\mathrm{BH}})}$.

As mentioned above, BH controls the false discovery rate defined as

$$\mathrm{FDR}(n) \equiv \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathbb{E}_{\theta}\left(\frac{V^{\theta}(n)}{R(n)\vee 1}\right)\,, \tag{3}$$

where $R(n)$ the number of rejected hypotheses. Note that BH requires the knowledge of *all* p-values to determine the significance level for testing the hypotheses. Hence, it does not address the online scenario.

In this paper, we study methods for *online* control of false discovery rate. Namely, we consider a sequence of hypotheses $H_1, H_2, H_3, \dots$ that arrive sequentially in a stream, with corresponding p-values p_1, p_2, $\dots$. We aim at developing a testing mechanism that ensures false discovery rate remains below a pre-assigned level α. A testing procedure provides a sequence of significance levels α_i, with decision rule:

$$R_i = \begin{cases} 1 & \text{if } p_i \le \alpha_i \qquad \text{(reject } H_i\text{)}, \\ 0 & \text{otherwise} \qquad \text{(accept } H_i\text{)}. \end{cases} \tag{4}$$

In *online* testing, we require significance levels to be functions of prior outcomes:

$$\alpha_i = \alpha_i(R_1, R_2, \dots, R_{i-1})\,. \tag{5}$$

Foster and Stine [FS07] introduced the above setting and proposed a class of procedures named *alpha-investing rules.* Alpha-investing starts with an initial wealth, at most α, of allowable false discovery rate. The wealth is spent for testing different hypotheses. Each time a discovery occurs, the alpha-investing procedure earns a contribution toward its wealth to use for further tests. Foster and Stine [FS07] proved that alpha-investing rules control a modified metric known as mFDR, i.e. the ratio of the expected number of false discoveries to the expected number of discoveries. As illustrated in Appendix A, mFDR and FDR can be very different in situations with high variability. While FDR is the expected proportion of false discoveries, mFDR is the ratio of two expectations and hence is not directly related to any single sequence quantity.

Several recent papers [LTTT14, GWCT15, LB16] consider a 'sequential hypothesis testing' problem that arises in connection with sparse linear regression. Let us emphasize that the problem treated in [LTTT14, GWCT15] is substantially different from the one analyzed here. For instance, as discussed in Section 8, the methods of [GWCT15] achieve vanishingly small statistical power for the present problem.

1.1 Contributions

In this paper, we study a class of procedures that are known as *generalized alpha-investing*, and were first introduced by Aharoni and Rosset in [AR14]. As in alpha-investing [FS07], generalized alpha-investing makes use of a potential sequence (wealth) that increases every time a null hypothesis is rejected, and decreases otherwise. However: (*i*) The pay-off and pay-out functions are general functions of past history; (*ii*) The pay-out is not tightly determined by the testing level α_i. This additional freedom allows to construct interesting new rules.

The contributions of this paper are summarized as follows.

Online control of FDR. We prove that generalized alpha-investing rules control FDR, under the assumption of independent p-values. To the best of our knowledge, this is the first work[1] that guarantees online control of FDR.

[1] Special cases were presented in our earlier technical report [JM15].

Online control of FDR for dependent p-values. Dependencies among p-values can arise for multiple reasons. For instance the same data can be re-used to test a new hypothesis, or the choice of a new hypothesis can depend on the past outcomes. We present a general upper bound on the FDR for dependent p-values under generalized alpha-investing.

False discovery exceedance. FDR can be viewed as the expectation of false discovery proportion (FDP). In some cases, FDP may not be well represented by its expectation, e.g., when the number of discoveries is small. In these cases, FDP might be sizably larger than its expectation with significant probability. In order to provide tighter control, we develop bounds on the false discovery exceedance (FDX), i.e. on the tail probability of FDP.

Statistical power. In order to compare different procedures, we develop lower bounds on fraction of non-null hypotheses that are discovered (statistical power), under a mixture model where each null hypothesis is false with probability π_1, for a fixed arbitrary π_1.

We focus in particular on a concrete example of generalized alpha-investing rule (called LORD below) that we consider particularly compelling. We use our lower bound to guide the choice of parameters for this rule.

Numerical Validation. We validate our procedures on synthetic and real data in Sections 5 and 7, showing that they control FDR and mFDR in an online setting. We further compare them with BH and Bonferroni procedures. We observe that generalized alpha investing procedures have statistical power comparable to BH, while satisfying the online constraint.

1.2 Notations

Throughout the paper, we typically use upper case symbols (e.g. $X, Y, Z, \dots$) to denote random variables, and lower case symbols for deterministic values (e.g. $x, y, z, \dots$). Vectors are denoted by boldface, e.g. $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \dots$ for random vectors, and $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \dots$ for deterministic vectors. Given a vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$, we use $\boldsymbol{X}_i^j = (X_i, X_{i+1}, \dots, X_j)$ to denote the sub-vector with indices between i and j. We will often consider sequences indexed by the same 'time index' as for the hypotheses $\{H_1, H_2, H_3, \dots\}$. Given such a sequence $(X_i)_{i\in\mathbb{N}}$, we denote by $X(n) \equiv \sum_{i=1}^n X_i$ its partial sums.

We denote the standard Gaussian density by $\phi(x) = e^{-x^2/2}/\sqrt{2\pi}$, and the Gaussian distribution function by $\Phi(x) = \int_{-\infty}^x \phi(t)\,\mathrm{d}t$. We use the standard big-O notation. In particular $f(n) = O(g(n))$ as $n \to \infty$ if there exists a constant $C > 0$ such that $|f(n)| \le C\, g(n)$ for all n large enough. We also use $\sim$ to denote asymptotic equality, i.e. $f(n) \sim g(n)$ as $n \to \infty$, means $\lim_{n\to\infty} f(n)/g(n) = 1$. We further use $\asymp$ for equality up to constants, i.e. if $f(n) = \Theta(g(n))$, then there exist constants $C_1, C_2 > 0$ such that $C_1|g(n)| \le f(n) \le C_2|g(n)|$ for all n large enough.

2 Generalized alpha-investing

In this section we define generalized alpha investing rules, and provide some concrete examples. Our definitions and notations follow the paper of Aharoni and Rosset that first introduced generalized alpha-investing [AR14].

2.1 Definitions

Given a sequence of input p-values $(p_1, p_2, p_3, \dots)$, a *generalized alpha-investing* rule generates a sequence of decisions $(R_1, R_2, R_3, \dots)$ (here $R_j \in \{0, 1\}$ and $R_j = 1$ is to be interpreted as rejection of null hypothesis H_j) by using test levels $(\alpha_1, \alpha_2, \alpha_3, \dots)$. After each decision j, the rule updates a potential function $W(j)$ as follows:

- If hypothesis j is accepted, then the potential function is decreased by a pay-out φ_j.
- If hypothesis j is rejected, then the potential is increased by an amount $\psi_j - \varphi_j$.

In other words, the pay-out φ_j is the amount paid for testing a new hypothesis, and the pay-off ψ_j is the amount earned if a discovery is made at that step.

Formally, a generalized alpha-investing rule is specified by three (sequences of) functions $\alpha_j, \varphi_j, \psi_j : \{0, 1\}^{j-1} \to \mathbb{R}_{\geq 0}$, determining test levels, pay-out and pay-off. Decisions are taken by testing at level α_j

$$R_j = \begin{cases} 1, & \text{if } p_j \leq \alpha_j = \alpha_j(R_1, \dots, R_{j-1}), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The potential function is updated via:

$$W(0) = w_0 \,, \tag{7}$$

$$W(j) = W(j-1) - \varphi_j(\boldsymbol{R}_1^{j-1}) + R_j \, \psi_j(\boldsymbol{R}_1^{j-1}) \,, \tag{8}$$

with $w_0 \geq 0$ an initial condition. Notice in particular that $W(j)$ is a function of $(R_1, \dots, R_j)$

A valid generalized alpha-investing rule is required to satisfy the following conditions, for a constant $b_0 > 0$:

G1. For all $j \in \mathbb{N}$ and all $\boldsymbol{R}_1^{j-1} \in \{0, 1\}^{j-1}$, letting $\psi_j = \psi_j(\boldsymbol{R}_1^{j-1})$, $\varphi_j = \varphi_j(\boldsymbol{R}_1^{j-1})$, $\alpha_j = \alpha_j(\boldsymbol{R}_1^{j-1})$, we have

$$\psi_j \leq \varphi_j + b_0 \,, \tag{9}$$

$$\psi_j \leq \frac{\varphi_j}{\alpha_j} + b_0 - 1 \,, \tag{10}$$

$$\varphi_j \leq W(j-1) \,. \tag{11}$$

G2. For all $j \in \mathbb{N}$, and all $\boldsymbol{R}_1^{j-1} \in \{0, 1\}^{j-1}$ such that $W(j-1) = 0$, we have $\alpha_j = 0$.

Notice that Condition (11) and G2 are well posed since $W(j-1)$, φ_j and α_j are functions of $\boldsymbol{R}_1^{j-1}$. Further, because of (11), the function $W(j)$ remains non-negative for all $j \in \mathbb{N}$.

Throughout, we shall denote by $\mathcal{F}_j$ the σ-algebra generated by the random variables $\{R_1, \dots, R_j\}$.

For $\boldsymbol{x}$, $\boldsymbol{y} \in \{0, 1\}^n$, we write $\boldsymbol{x} \preceq \boldsymbol{y}$ if $x_j \leq y_j$ for all $j \in \{1, \dots, n\}$. We say that an online rule is *monotone* if the functions α_j are monotone non-decreasing with respect to this partial ordering (i.e. if $\boldsymbol{x} \preceq \boldsymbol{y}$ implies $\alpha_j(\boldsymbol{x}) \leq \alpha_j(\boldsymbol{y})$).

Remark 2.1. Our notation differs from [AR14] in one point, namely we use w_0 for the initial potential (which is denoted by $\alpha\eta$ in [AR14]) and b_0 for the constant appearing in Eqs. (9), (10) (which is denoted by α in [AR14]). We prefer to reserve α for the FDR level[2].

[2] The use of η in [AR14] was related to control of mFDR$_\eta$ in that paper.

Remark 2.2. In a generalized alpha investing rule, as we reject more hypotheses the potential $W(j)$ increases and hence we can use large test levels α_j. In other words, the burden of proof decreases as we reject more hypotheses. This is similar to the BH rule, where the most significant p-values is compared to a Bonferroni cutoff, the second most significant to twice this cutoff and so on. However, it is worth noting that in BH the test levels α_j are at most α, while in a generalized alpha-investing test rule, test levels could possibly grow larger than α and in this sense, test levels in generalized alpha-investing are more flexible than in BH.

2.2 Examples

Generalized α-investing rules comprise a large variety of online hypothesis testing methods. We next describe some specific subclasses that are useful for designing specific procedures.

2.2.1 Alpha Investing

Alpha investing, introduced by Foster and Stine [FS07], is a special case of generalized alpha-investing rule. In this case the potential is decreased by $\alpha_j/(1-\alpha_j)$ if hypothesis H_j is not rejected, and increased by a fixed amount b_0 if it is rejected. In formula, the potential evolves according to

$$W(j) = W(j-1) - (1-R_j)\frac{\alpha_j}{1-\alpha_j} + R_j b_0\,. \tag{12}$$

This fits the above framework by defining $\varphi_j = \alpha_j/(1-\alpha_j)$ and $\psi_j = b_0 + \alpha_j/(1-\alpha_j)$. Note that this rule depends on the choice of the test levels α_j, and of the parameter b_0. The test levels α_j can be chosen arbitrarily, provided that they satisfy condition (11), which is equivalent to $\alpha_j/(1-\alpha_j) \le W(j-1)$.

2.2.2 Alpha Spending with Rewards

Alpha spending with rewards was introduced in [AR14], as a special sub-class of generalized alpha investing rules, which are convenient for some specific applications.

In this case, test levels are chosen to be proportional to the pay-out function, $\alpha_j = \varphi_j/\kappa$, with a proportionality coefficient κ. Conditions (9) and (10) coincide with[3]

$$0 \le \psi_j \le \min\Big(\kappa\alpha_j + b_0, \kappa - 1 + b_0\Big)\,. \tag{13}$$

The choice of penalties φ_j is arbitrary as long as constraint (11) is satisfied. For instance, [AR14] uses $\varphi_j = c_1 W(j-1)$ with $c_1 \in (0,1)$.

2.2.3 LORD

As a running example, we shall use a simple procedure that we term LORD , for Levels based On Recent Discovery. LORD is easily seen to be a special case of alpha spending with rewards, for $\kappa = 1$.

[3] Note that [AR14] rescales the potential function by κ, and hence the condition on ψ_j is also rescaled.

In LORD , the significance levels α_i depend on the past only through the time of the last discovery, and the wealth accumulated at that time. Concretely, choose any sequence of non-negative numbers $\boldsymbol{\gamma} = (\gamma_i)_{i\in\mathbb{N}}$, which is monotone non-increasing (i.e. for $i \le j$ we have $\gamma_i \ge \gamma_j$) and such that $\sum_{i=1}^{\infty} \gamma_i = 1$. We refer to Section 4 for concrete choices of this sequence.

At each time i, let τ_i be the last time a discovery was made before i. We then use a fraction $\gamma_{i-\tau_i}$ of the wealth available at time τ_i. If a discovery is made, we add an amount b_0 to the current wealth. Formally, we set

$$W(0) = w_0\,, \tag{14}$$

$$\varphi_i = \alpha_i = \gamma_{i-\tau_i}\, W(\tau_i)\,, \tag{15}$$

$$\psi_i = b_0\,, \tag{16}$$

$$\tau_i \equiv \max\Big\{\ell \in \{1,\dots,i-1\}\ :\ R_\ell = 1\Big\}\,, \tag{17}$$

where, by convention, $\tau_1 = 0$, and $\{W(j)\}_{j\ge 0}$ is defined recursively via Eq. (8) . Note that τ_i is measurable on $\mathcal{F}_{i-1}$, and hence φ_i, ψ_i are functions of $\boldsymbol{R}_1^{i-1}$ as claimed, while $W(i)$ is a function of $\boldsymbol{R}_1^{i}$. This is obviously an online multiple hypothesis testing procedure, and is monotone by the monotonicity of γ_i.

Condition G1 is easily verified. Specifically, Equations (9) and (10) hold because $\alpha_j = \varphi_j$ and $\psi_j = b_0$. Equation (11) stands since

$$W(j-1) = W(\tau_j) - \sum_{i=\tau_j+1}^{j-1} \varphi_i = W(\tau_j)\Big\{1 - \sum_{k=1}^{j-1-\tau_j} \gamma_k\Big\} \tag{18}$$

$$\ge W(\tau_j)\,\gamma_{j-\tau_j} = \varphi_j\,. \tag{19}$$

Finally, condition G2 follows easily because $W(i) = 0$ implies $\alpha_j = 0$ and $W(j) = 0$ for all $j \ge i$.

We conclude that LORD is a monotone generalized alpha-investing rule.

3 Control of false discovery rate

3.1 FDR control for independent test statistics

As already mentioned, we are interested in testing a –possibly infinite– sequence of null hypotheses $\mathcal{H} = (H_i)_{i\in\mathbb{N}}$. The set of first n hypotheses will be denoted by $\mathcal{H}(n) = (H_i)_{1\le i\le n}$. Without loss of generality, we assume H_i concerns the value of a parameter θ_i, with $H_i = \{\theta_i = 0\}$. Rejecting the null hypothesis H_i can be interpreted as θ_i being significantly non-zero. We will denote by Θ the set of possible values for the parameters θ_i, and by $\boldsymbol{\Theta} = \Theta^{\mathbb{N}}$ the space of possible values of the sequence $\boldsymbol{\theta} = (\theta_i)_{i\in\mathbb{N}}$

Under the null hypothesis $H_i : \theta_i = 0$, the corresponding p-value is uniformly random in $[0,1]$:

$$p_i \sim \mathsf{Unif}([0,1])\,. \tag{20}$$

Recall that R_i is the indicator that a discovery is made at time i, and $R(n) = \sum_{i=1}^{n} R_i$ the total number of discoveries up to time n. Analogously, let $V_i^{\boldsymbol{\theta}}$ be the indicator that a false discovery occurs at time i and $V^{\theta}(n) = \sum_{i=1}^{n} V_i^{\theta}$ the total number of false discovery up to time n. Throughout the paper, superscript θ is used to distinguish unobservable variables such as $V^\theta(n)$, from statistics such as $R(n)$. However, we drop the superscript when it is clear from the context.

There are various criteria of interest for multiple testing methods. We will mostly focus on the *false discovery rate (FDR)* [BH95], and we repeat its definition here for the reader's convenience. We first define the *false discovery proportion* (FDP) as follows. For $n \ge 1$,

$$\mathrm{FDP}^\theta(n) \equiv \frac{V^\theta(n)}{R(n) \vee 1} \, . \tag{21}$$

The false discovery rate is defined as

$$\mathrm{FDR}(n) \equiv \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_\theta\Big(\mathrm{FDP}^\theta(n)\Big) \, . \tag{22}$$

Our first result establishes FDR control for all monotone generalized alpha-investing procedures. Its proof is presented in Appendix B.

Theorem 3.1. *Assume the p-values $(p_i)_{i\in\mathbb{N}}$ to be independent. Then, for any monotone generalized alpha-investing rule with $w_0 + b_0 \le \alpha$, we have*

$$\sup_n \mathrm{FDR}(n) \le \alpha \, . \tag{23}$$

The same holds if only the p-values corresponding to true nulls are mutually independent, and independent from the non-null p-values.

Remark 3.2. Appendix B proves a somewhat stronger version of Theorem 3.1, namely $\mathrm{FDR}(n) \le b_0\mathbb{E}\{R(n)/(R(n) \vee 1)\} + w_0\mathbb{E}\{1/(R(n) \vee 1)\}$. In particular, $\mathrm{FDR}(n) \lesssim b_0$ when the total number of discoveries $R(n)$ is large, with high probability. This is the case –for instance– when the hypotheses to be tested comprise a large number of 'strong signals' (even if these form a small proportion of the total number of hypotheses).

Another possible strengthening of Theorem 3.1 is obtained by considering a new metric, that we call $\mathrm{sFDR}_\eta(n)$ (for smoothed FDR):[4]

$$\mathrm{sFDR}_\eta(n) \equiv \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}\Big\{\frac{V^\theta(n)}{R(n) + \eta}\Big\} \, . \tag{24}$$

The following theorem bounds $\mathrm{sFDR}_{w_0/b_0}(n)$ for monotone generalized alpha-investing rules.

Theorem 3.3. *Under the assumptions of Theorem 3.1, for any $w_0, b_0 > 0$, we have*

$$\sup_n \mathrm{sFDR}_{w_0/b_0}(n) \le b_0 \, . \tag{25}$$

Note that Eq. (25) implies (23) by using $R(n) + (w_0/b_0) \le (b_0 + w_0)R(n)/b_0$ for $R(n) \ge 1$. Also, $\mathbb{E}\big\{V^\theta(n)/(R(n) + (w_0/b_0))\big\} \approx \mathrm{FDR}(n)$ if $R(n)$ is large with high probability.

Let us emphasize that the guarantee in Theorem 3.3 is different from the one in [FS07, AR14], which instead use $\mathrm{mFDR}_\eta(n) \equiv \mathbb{E}\{V^\theta(n)\}/(\mathbb{E}\{R(n)\} + \eta)$ (the latter is a ratio of expectations instead of expectation of the ratio, and, as mentioned, does not correspond to a single-sequence property).

[4] Some authors [BC16] refer to this quantity as "modified FDR". We will not follow this terminology since its acronym (mFDR) gets confused with "marginal FDR" [FS07, AR14].

Remark 3.4. In Appendix C we show that Theorems 3.1 and 3.3 cannot be substantially improved, unless specific restrictions are imposed on the generalized alpha-investing rule. In particular, we prove that there exist generalized alpha investing rules for which $\liminf_{n\to\infty} \mathrm{FDR}(n) \ge b_0$, and $\lim_{n\to\infty} \mathrm{sFDR}_{w_0/b_0} = b_0$.

3.2 FDR control for dependent test statistics

In some applications, the assumption of independent p-values is not warranted. This is the case –for instance– of multiple related hypotheses being tested on the same experimental data. Benjamini and Yekutieli [BY01] introduced a property called *positive regression dependency from a subset* I_0 (PRDS on I_0) to capture a positive dependency structure among the test statistics. They showed that if the joint distribution of the test statistics is PRDS on the subset of test statistics corresponding to true null hypotheses, then BH controls FDR. (See Theorem 1.3 in [BY01].) Further, they proved that BH controls FDR under general dependency if its threshold is adjusted by replacing α with $\alpha/(\sum_{i=1}^{N}\frac{1}{i})$ in equation (2).

Our next result establishes an upper bound on the FDR of generalized alpha-investing rules, under general p-values dependencies. For a given generalized alpha-investing rule, let $\mathcal{R}_i \equiv \{\boldsymbol{r}_1^i \in \{0,1\}^i : \mathbb{P}(\boldsymbol{R}_1^i = \boldsymbol{r}_1^i) > 0\}$, the set of decision sequences that have non-zero probability.

Definition 3.5. *An* index sequence *is a sequence of deterministic functions $\mathcal{I} = (\mathcal{I}_i)_{i\in\mathbb{N}}$ with $\mathcal{I}_i : \{0,1\}^i \to \mathbb{R}_{\ge 0}$. For an index sequence $\mathcal{I}$, let*

$$R_i^{\mathrm{L}}(s) \equiv \min_{\boldsymbol{r}_1^{i-1}\in\mathcal{R}_{i-1}} \Big\{ \sum_{j=1}^{i-1} r_j \,:\, \mathcal{I}_{i-1}(\boldsymbol{r}_1^{i-1}) \ge s \Big\}\,, \tag{26}$$

$$\mathcal{I}_{\min}(i) \equiv \min_{\boldsymbol{r}_1^{i}\in\mathcal{R}_{i}} \mathcal{I}_i(\boldsymbol{r}_1^i)\,,\;\; \mathcal{I}_{\max}(i) \equiv \max_{\boldsymbol{r}_1^{i}\in\mathcal{R}_{i}} \mathcal{I}_i(\boldsymbol{r}_1^i)\,. \tag{27}$$

As concrete examples of the last definition, for a generalized alpha-investing rule, the current potentials $\{W(i)\}_{i\in\mathbb{N}}$, potentials at the last rejection $\{W(\tau_i)\}_{i\in\mathbb{N}}$ and total number of rejections $\{R(i)\}_{i\in\mathbb{N}}$ are index sequences.

Theorem 3.6. *Consider a generalized alpha-investing rule and assume that the test level α_j is determined based on index function $\mathcal{I}_{j-1}$. Namely, for each $j \in \mathbb{N}$ there exists a function $g_j : \mathbb{R}_{\ge 0} \to [0,1]$ such that $\alpha_j = g_j(\mathcal{I}_{j-1}(\boldsymbol{R}_1^{j-1}))$. Further, assume $g_j(\,\cdot\,)$ to be nondecreasing and weakly differentiable with weak derivative $\dot{g}_j(s)$.*

Then, the following upper bound holds for general dependencies among p-values:

$$\mathrm{FDR}(n) \le \sum_{i=1}^{n} \left\{ g_i(\mathcal{I}_{\min}(i-1)) + \int_{\mathcal{I}_{\min}(i-1)}^{\mathcal{I}_{\max}(i-1)} \frac{\dot{g}_i(s)}{R_i^{\mathrm{L}}(s)+1} \mathrm{d}s \right\}. \tag{28}$$

The proof of this theorem is presented in Appendix D.

Example 3.7 (FDR control for dependent test statistics via modified Lord)**.** We can modify Lord as to achieve FDR control even under dependent test statistics. As before, we let $\psi_i = b_0$. However, we fix a sequence $\boldsymbol{\xi} = (\xi_i)_{i\in\mathbb{N}}$, $\xi_i \ge 0$, and set test levels according to rule $\alpha_i = \varphi_i = \xi_i\, W(\tau_i)$. In other words, compared with the original Lord procedure, we discount the capital accumulated at the last discovery as a function of the number of hypotheses tested so far, rather than the number of hypotheses tested since the last discovery.

This rule satisfies the assumptions of Theorem 3.6, with index sequence $\mathcal{I}_{i-1} = W(\tau_i)$ and $g_i(s) = \xi_i s$. Further, $\mathcal{I}_{\min}(0) = w_0$, $\mathcal{I}_{\min}(i-1) = b_0$ for $i \geq 2$, and $\mathcal{I}_{\max}(i-1) \leq w_0 + b_0(i-1)$, and $R_i^{\mathrm{L}}(s) \geq (\frac{s-w_0}{b_0})_+$. Substituting in Eq. (28), we obtain, assuming $w_0 \leq b_0$

$$\begin{aligned}
\mathrm{FDR}(n) &\leq w_0\xi_1 + \sum_{i=2}^{n} \left(b_0\xi_i + \int_{b_0}^{w_0+b_0(i-1)} \frac{b_0\xi_i}{s - w_0 + b_0}\,\mathrm{d}s \right) \\
&\leq w_0\xi_1 + \sum_{i=2}^{n} b_0\xi_i(1 + \log(i)) \\
&\leq \sum_{i=1}^{n} b_0\xi_i(1 + \log(i))\,.
\end{aligned}$$

Therefore, this rule controls FDR below level α under general dependency structure, if we choose coefficients $(\xi_i)_{i\in\mathbb{N}}$ such that $\sum_{i=1}^{\infty} \xi_i(1 + \log(i)) \leq \alpha/b_0$.

4 Statistical power

The class of generalized alpha-investing rules is quite broad. In order to compare different approaches, it is important to estimate their statistical power.

Here, we consider a mixture model wherein each null hypothesis is false with probability π_1 independently of other hypotheses, and the p-values corresponding to different hypotheses are mutually independent. Under the null hypothesis H_i, we have p_i uniformly distributed in $[0, 1]$ and under its alternative, p_i is generated according to a distribution whose c.d.f is denoted by F. We let $G(x) = \pi_0\, x + \pi_1\, F(x)$, with $\pi_0 + \pi_1 = 1$, be the marginal distribution of the p-values. For presentation clarity, we assume that $F(x)$ is continuous.

While the mixture model is admittedly idealized, it offers a natural ground to compare online procedures to offline procedures. Indeed, online approaches are naturally favored if the true non-nulls arise at the beginning of the sequence of hypotheses, and naturally unfavored if they only appear later. On the other hand, if the p-values can be processed offline, we can always apply an online rule after a random re-ordering of the hypotheses. By exchangeability, we expect the performance to be similar to the ones in the mixture model.

The next theorem lower bounds the statistical power of LORD under the mixture model.

Theorem 4.1. *Consider the mixture model with $G(x)$ denoting the marginal distribution of p-values. Further, let $\Omega_0(n)$ (and its complement $\Omega_0^c(n)$) be the subset of true nulls (non-nulls), among the first n hypotheses. Then, the average power of* LORD *rule is almost surely bounded as follows:*

$$\liminf_{n\to\infty} \frac{1}{|\Omega_0^c(n)|} \sum_{i\in\Omega_0^c(n)} R_i \geq \Big(\sum_{m=1}^{\infty} \prod_{\ell=1}^{m} \big(1 - G(b_0\gamma_\ell)\big) \Big)^{-1}. \tag{29}$$

Theorem 4.1 is proved in Appendix F. The lower bound is in fact the exact power for a slightly weaker rule that resets the potential at level b_0 after each discovery (in other words, Eq. (15) is replaced by $\varphi_i = \gamma_{i-\tau_i} b_0$). This procedure is weaker only when multiple discoveries are made in a short interval of time. Hence, the above bound is expected to be accurate when π_1 is small, and discoveries are rare.

Recall that in LORD , parameters $\boldsymbol{\gamma} = (\gamma_\ell)_{\ell=1}^\infty$ can be any sequence of non-negative, monotone non-increasing numbers that sums up to one. This leaves a great extent of flexibility in choosing γ. The above lower bound on statistical power under the mixture model provides useful insight on what are good choices of $\boldsymbol{\gamma}$.

We first simplify the lower bound further. We notice that $\prod_{\ell=1}^m \big(1 - G(b_0\gamma_\ell)\big) \le \exp\big(-\sum_{\ell=1}^m G(b_0\gamma_\ell)\big)$. Further, by the monotonicity property of $\boldsymbol{\gamma}$, we have $G(b_0\gamma_\ell) \ge G(b_0\gamma_m)$ for $\ell \le m$. Thus,

$$\lim_{n\to\infty} \frac{1}{|\Omega_0^c(n)|} \sum_{i\in\Omega_0^c(n)} R_i \ge \mathcal{A}(G,\boldsymbol{\gamma})\,, \quad \mathcal{A}(G,\boldsymbol{\gamma}) = \Big(\sum_{m=1}^\infty e^{-mG(b_0\gamma_m)}\Big)^{-1}\,. \tag{30}$$

In order to choose $\boldsymbol{\gamma}$, we use the lower bound $\mathcal{A}(G,\boldsymbol{\gamma})$ as a surrogate objective function. We let $\boldsymbol{\gamma}^{\rm opt}$ be the sequence that maximizes $\mathcal{A}(G,\boldsymbol{\gamma})$. The following proposition characterizes the asymptotic behavior of $\boldsymbol{\gamma}^{\rm opt}$.

Proposition 4.2. *Let $\boldsymbol{\gamma}^{\rm opt}$ be the sequence that maximizes $\mathcal{A}(G,\boldsymbol{\gamma})$ under the constraint $\sum_{\ell=1}^\infty \gamma_m = 1$. Further suppose that $F(x)$ is concave and differentiable on an interval $[0,x_0)$ for some $x_0 \in (0,1)$. Then there is a constant $\eta = \eta(G,\pi_1)$ independent of m such that, for all m large enough, the following holds true:*

$$\frac{1}{b_0}G^{-1}\Big(\frac{1}{m}\log\Big(\frac{m(1-\pi_1)}{\eta}\Big)\Big) \le \gamma_m^{\rm opt} \le \frac{1}{b_0}G^{-1}\Big(\frac{2}{m}\log\Big(\frac{1}{\eta G^{-1}(1/m)}\Big)\Big)\,. \tag{31}$$

The proof of Proposition 4.2 is deferred to Appendix G.

The concavity assumption of $F(x)$ requires the density of non-null p-values (i.e., $F'(x)$) to be non-increasing in a neighborhood $[0,x_0)$. This is a reasonable assumption because significant p-values are generically small and the assumption states that, in a neighborhood of zero, smaller values have higher density than larger values.

Example 4.3. Suppose that non-null p-values are generated as per Beta density with parameters $a,b > 0$. Then $F(x) = I_x(a,b)$ where $I_x(a,b) = (\int_0^x t^{a-1}(1-t)^{b-1}{\rm d}t)/B(a,b)$ is the regularized incomplete Beta function and $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}{\rm d}t$ denotes the Beta function. It is easy to see that for $a < 1$ and $b \ge 1$, $F(x)$ is concave. Moreover, $\lim_{x\to 0} x^a/F(x) = aB(a,b)$. Hence, for $a < 1$, we get $G(x) = \pi_1 F(x) + (1-\pi_1)x \asymp x^a$, up to constant factor that depends on a, b, π_1. Applying Proposition 4.2, we obtain

$$\gamma_m \asymp \Big(\frac{1}{m}\log m\Big)^{1/a}. \tag{32}$$

When a is small, the beta density $F'(x)$ decreases very rapidly with x, and thus the non-null p-values are likely to be very small. It follows from Eq. (32) that γ_m also decreases rapidly in this case. This is intuitively justified: when the non-null p-values are typically small, small test levels are adequate to reject the true non-nulls and not to waste the α-wealth. On the other hand, when a grows, the range of significant p-values becomes broader and the coefficients γ_m decay more slowly.

Example 4.4. (Mixture of Gaussians) Suppose we are getting samples $Z_j \sim \mathsf{N}(\theta_j, 1)$ and we want to test null hypotheses $H_j:\ \theta_j = 0$ versus alternative $\theta_j = \mu$. In this case, two-sided p-values are given by $p_j = 2\Phi(-|Z_j|)$ and hence

$$\begin{aligned} F(x) &= \mathbb{P}_1\big(|Z_j| \ge \Phi^{-1}(1-x/2)\big) \\ &= \Phi(-\Phi^{-1}(1-x/2) - \mu) + \Phi(\mu - \Phi^{-1}(1-x/2))\,. \end{aligned} \tag{33}$$

Recall the following classical bound on the c.d.f of normal distribution for $t \ge 0$:

$$\frac{\phi(t)}{t}\Big(1-\frac{1}{t^2}\Big) \le \Phi(-t) \le \frac{\phi(t)}{t}\,. \tag{34}$$

Define $\xi(x) = \Phi^{-1}(1-x/2)$, and hence $x = 2\Phi(-\xi(x))$. A simple calculation shows that

$$\lim_{x\to 0} \xi(x)/\sqrt{2\log(1/x)} = 1\,. \tag{35}$$

Applying inequalities (34) and Eq. (35), simple calculus shows that, as $x \to 0$,

$$F(x) \sim \frac{1}{2}\, x\, e^{-\mu^2/2} e^{\mu\sqrt{2\log(1/x)}}\,. \tag{36}$$

Hence, for $G(x) = \pi_1 F(x) + (1-\pi_1)x$, we obtain

$$G(x) \sim \frac{\pi_1}{2}\, x\, e^{-\mu^2/2} e^{\mu\sqrt{2\log(1/x)}}\,. \tag{37}$$

Using Proposition 4.2, we obtain that for large enough m,

$$C_1 \frac{\log m}{m e^{\mu\sqrt{2\log m}}} \le \gamma_m^{\rm opt} \le C_2 \frac{\log m}{m e^{\mu\sqrt{2\log m}}}\,, \tag{38}$$

with C_1, C_2 constants depending only on μ, b_0. (In particular, we can take $C_1(\mu,\alpha) = 1.9\, b_0^{-1} e^{\mu^2/2}$, $C_2(\mu,\alpha) = 4.1\, b_0^{-1} e^{\mu^2/2}$.)

5 Numerical simulations

In this section we carry out some numerical experiments with synthetic data. For an application with real data, we refer to Section 7.

5.1 Comparison with off-line rules

In our first experiment, we consider hypotheses $\mathcal{H}(n) = (H_1, H_2, \dots, H_n)$ concerning the means of normal distributions. The null hypothesis is $H_j : \theta_j = 0$. We observe test statistics $Z_j = \theta_j + \varepsilon_j$, where ε_j are independent standard normal random variables. Therefore, one-sided p-values are given by $p_j = \Phi(-Z_j)$, and two sided p-values by $p_j = 2\Phi(-|Z_j|)$. Parameters θ_j are set according to a mixture model:

$$\theta_j \sim \begin{cases} 0 & \text{w.p.} \quad 1-\pi_1\,, \\ F_1 & \text{w.p.} \quad \pi_1\,. \end{cases} \tag{39}$$

In our experiment, we set $n = 3000$ and and use the following three choices of the non-null distribution:

Gaussian. In this case the alternative F_1 is $\mathsf{N}(0,\sigma^2)$ with $\sigma^2 = 2\log n$. This choice of σ produces parameters θ_j in the interesting regime in which they are detectable, but not easily so. In order to see this recall that, under the global null hypothesis, $Z_i \sim \mathsf{N}(0,1)$ and $\max_{i\in[n]} Z_i \sim \sqrt{2\log n}$ with high probability. Indeed $\sqrt{2\log n}$ is the minimax amplitude for estimation in the sparse Gaussian sequence model [DJ94, Joh94].

In this case we carry out two-sided hypothesis testing.

Exponential. In this case the alternative F_1 is exponential $\mathsf{Exp}(\lambda)$ with mean $\lambda^{-1} = \sqrt{2\log n}$. The rationale for this choice is the same given above.

The alternative is known to be non-negative, and hence we carry out one-sided hypothesis testing.

Simple. In this example, the non-nulls are constant and equal to $A = \sqrt{\log n}$. Again, we carry out one-sided tests in this case.

We consider three online testing rules, namely alpha investing (AI), LORD (a special case of alpha spending with rewards) and Bonferroni. For the case of simple alternatives, we also simulate the expected reward optimal (ERO) alpha investing rule introduced in [AR14] . We also consider the BH procedure. As emphasized already above BH is an offline testing rule: it has access to the number of hypotheses and p-values in advance, while the former algorithms receive p-values in an online manner, without knowing the total number of hypotheses. We use Storey's variant of BH rule, that is better suited to cases in which the fraction of non-nulls π_1 is not necessarily small [Sto02]. In all cases, we set as our objective to control FDR below $\alpha = 0.05$.

The different procedures are specified as follows:

Alpha Investing. We set test levels according to

$$\alpha_j = \frac{W(j)}{1 + j - \tau_j}\,, \tag{40}$$

where τ_j denotes the time of the most recent discovery before time j. This proposal was introduced by [FS07] and boosts statistical power in cases in which the non-null hypotheses appear in batches. We use parameters $w_0 = 0.005$ (for the initial potential), and $b_0 = \alpha - w_0 = 0.045$ (for the rewards). The rationale for this choice is that b_0 controls the evolution of the potential $W(n)$ for large n, while w_0 controls its initial value. Hence, the behavior of the resting rule for large n is mainly driven by b_0.

Note that, by [AR14, Corollary 2], this is an ERO alpha investing rule [5], under the Gaussian and exponential alternatives. In particular, for the exponential alternatives, it is expected reward optimal, i.e., it maximize the expected reward at the next step, cf. [AR14, Theorem 2].

ERO alpha-investing. For the case of simple alternative, the maximum power achievable at test i is $\rho_i = \Phi(\lambda + \Phi^{-1}(\alpha_i))$. In this case, we consider ERO alpha-investing [AR14] defined by $\varphi_i = (1/10) \cdot W(i-1)$, and with α_i, ψ_i given implicitly by the solution of $\varphi_i/\rho_i = \varphi_i/\alpha_i - 1$ and $\psi_i = \varphi_i/\alpha_i + b_0 - 1$. We use parameters $b_0 = 0.045$ and $w_0 = 0.005$.

LORD. We choose the sequence $\boldsymbol{\gamma} = (\gamma_m)_{m\in\mathbb{N}}$ as follows:

$$\gamma_m = C\,\frac{\log(m \vee 2)}{m e^{\sqrt{\log m}}}\,, \tag{41}$$

with C determined by the condition $\sum_{m=1}^{\infty} \gamma_m = 1$, which yields $C \approx 0.07720838$. This choice of $\boldsymbol{\gamma}$ is loosely motivated by Example 4.4. Notice however that we do not assume the data to be generated with the model treated in that example.

[5] Note that, since θ_j is unbounded under the alternative the maximal power is equal to one.

Also in this case, we use parameters $w_0 = 0.005$ (for the initial potential), and $b_0 = 0.045$ (for the rewards).

Bonferroni. We set the test levels as $\alpha_m = \gamma_m \alpha$, where the values of γ_m are set as per Equation (41), and therefore $\sum_{m=1}^{\infty} \alpha_m = \alpha$.

Storey. It is well known that the classical BH procedure satisfies FDR $\le \pi_0 \alpha$ where π_0 is the proportion of true nulls. A number of adaptive rules have been proposed that use a plug-in estimate of π_0 as a multiplicative correction in the BH procedure [Sto02, MR06, JC07, Jin08]. Following [BR09], the adaptive test thresholds are given by $\alpha H(\boldsymbol{p}) i/n$ (instead of $\alpha i/n$), where $H(\boldsymbol{p})$ is an estimate of π_0^{-1}, determined as a function of p-values, $\boldsymbol{p} = (p_1, \dots, p_n)$.

Here, we focus on Storey-λ estimator given by: [Sto02]

$$H(\boldsymbol{p}) = \frac{(1-\lambda)n}{\sum_{i=1}^{n} \mathbb{I}(p_i > \lambda) + 1} \,. \tag{42}$$

Storey's estimator is in general an overestimate of π_0^{-1}. A standard choice of $\lambda = 1/2$ is used in the SAM software [ST03]. Blanchard and Roquain [BR09] showed that the choice $\lambda = \alpha$ can have better properties under dependent p-values. In our simulations we tried both choices of λ.

Our empirical results are presented in Fig. 1. As we see all the rules control FDR below the nominal level $\alpha = 0.05$, as guaranteed by Theorem 3.1. While BH and the generalized alpha investing schemes (Lord , alpha-investing, ERO alpha investing) exploit most of the allowed amount of false discoveries, Bonferroni is clearly too conservative. A closer look reveals that the generalized alpha-investing schemes are somewhat more conservative that BH. Note however that the present simulations assume the non-nulls to arrive at random times, which is a more benign scenario than the one considered in Theorem 3.1.

In terms of power, generalized alpha-investing is comparable to BH for all the implementations considered (Lord , alpha investing and ERO alpha-investing). This suggests that under the present experiment setup, *online rules have power comparable to benchmark offline rules.* In particular, Lord appears particularly effective for small π_1, while standard alpha-investing suffers a loss of power for large π_1. This is related to the fact that $\varphi_j = \alpha_j/(1-\alpha_j)$ in this case. As a consequence the rule can effectively stop after a large number of discoveries, because α_j gets close to one.

Figure 2 showcases the FDR achieved by various rules as a function of α, for $\pi_1 = 0.2$ and exponential alternatives. For alpha-investing and Lord we use parameters $b_0 = 0.9\alpha$ and $w_0 = 0.1\alpha$. The generalized alpha-investing rules under consideration have FDR below the nominal α, and track it fairly closely. The gap is partly due to the fact that, for large number of discoveries, the FDR of generalized alpha investing rules is closer to b_0 than to $\alpha = b_0 + w_0$, cf. Remark 3.2.

5.2 The effect of ordering

By definition, the BH rule is insensitive to the order in which the hypotheses are presented. On the contrary, the outcome of online testing rules depends on this ordering. This is a weakness, because the ordering of hypotheses can be adversarial, leading to a loss of power, but also a strength. Indeed, in some applications, hypotheses which are tested early are the most likely to be non-null. In these cases, we expect generalized alpha-investing procedures to be potentially *more powerful than benchmark offline rules* as BH.

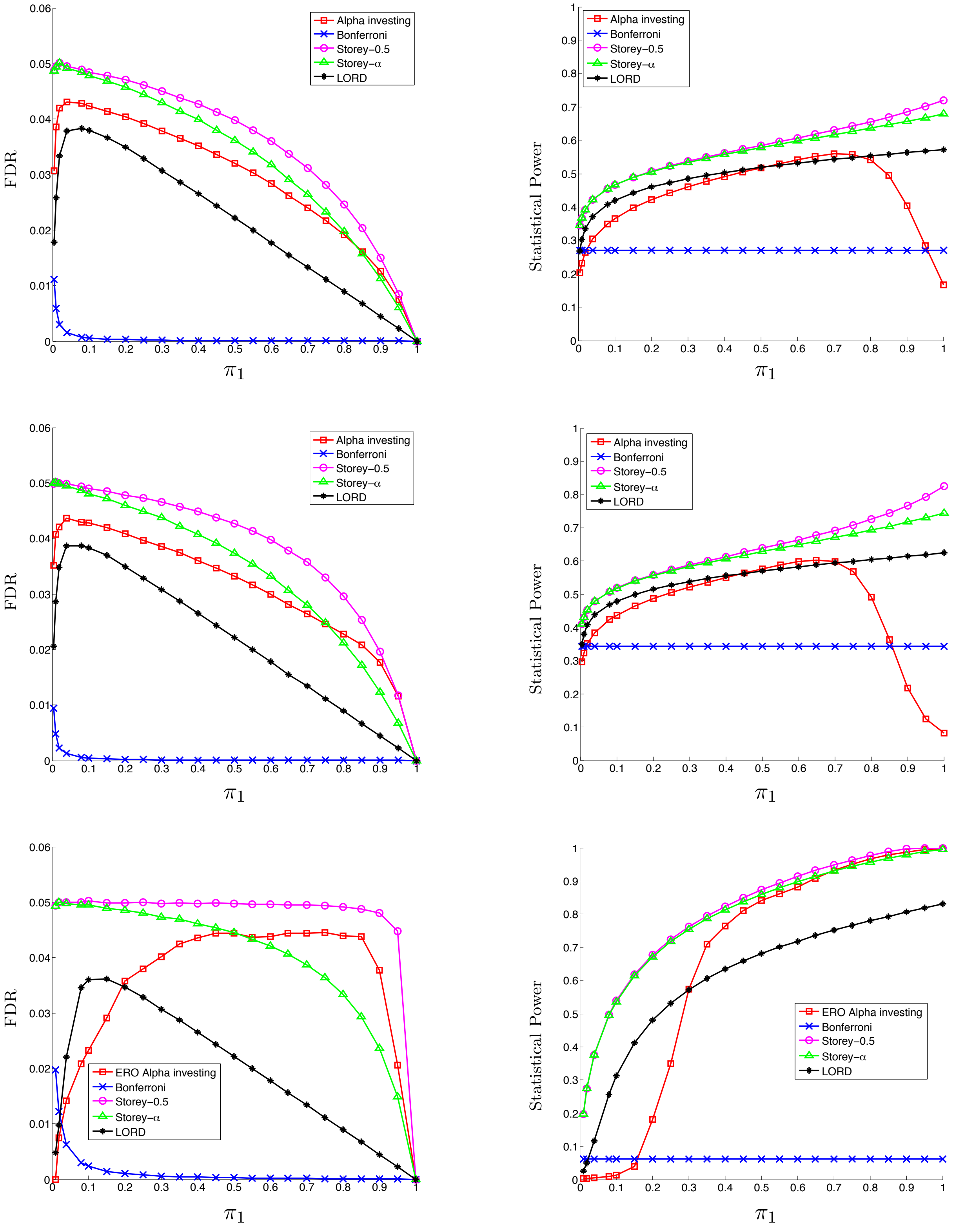


Figure 1: FDR and statistical power versus fraction of non-null hypotheses π_1 for setup described in Section 5 with $\alpha = 0.05$. The three rows correspond to Gaussian, exponential, and simple alternatives (from top to bottom). FDR and power are computed by averaging over 20,000 independent trials (for Gaussian and exponential alternatives) or 500 trials (for simple alternatives). Here hypotheses are considered in random order of arrival.

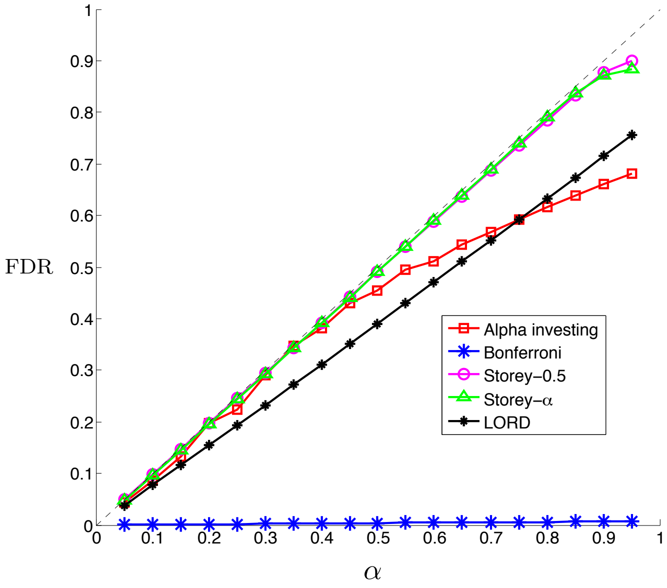


Figure 2: FDR achieved by various methods compared to the target FDR α as α varies. Here we use $n = 3000$ hypotheses with a proportion $\pi_1 = 0.2$ of non-nulls and exponential alternatives. The FDR is estimated by averaging over $20,000$ independent trials.

For instance, Li and Barber [LB16] analyze a drug-response dataset proceeding in two steps. First, a family of hypotheses (gene expression levels) are ordered using side information, and then a multiple hypothesis testing procedure is applied to the ordered data[6].

In order to explore the effect of a favorable ordering of the hypotheses, we reconsider the exponential model in the previous section, and simulate a case in which side information is available. For each trial, we generate the mean $(\theta_j)_{1\le j\le n}$, and two independent sets of observations $Z_j = \theta_j + \varepsilon_j$, $Z'_j = \theta_j + \varepsilon'_j$, with $\varepsilon_j \sim \mathsf{N}(0,1)$, $\varepsilon'_j \sim \mathsf{N}(0,\sigma^2)$ independent. We then compute the corresponding (one-sided) p-values $(p_j)_{1\le j\le n}$, $(p'_j)_{1\le j\le n}$. We use the p-values $(p'_j)_{1\le j\le n}$ to order the hypotheses[7] (in such a way that these p-values are increasing along the ordering). We then use the other set of p-values $(p_j)_{1\le j\le n}$ to test the null hypotheses $H_{j,0} : \theta_j = 0$ along this ordering.

Let us emphasize that, for this simulation, better statistical power would be achieved if we computed a single p-value p_j by processing jointly Z_j and Z'_j. However, in real applications, the two sources of information are heterogenous and this joint processing is not warranted, see [LB16] for a discussion of this point.

Figure 3 reports the FDR and statistical power in this setting. We used LORD with parameters $(\gamma_m)_{m\ge 1}$ given by Eq. (41), and simulated two noise levels for the side information: $\sigma^2 = 1$ (noisy ordering information) and $\sigma^2 = 1/2$ (less noisy ordering). As expected, with a favorable ordering the FDR decreases significantly. The statistical power increases as long as the fraction of non-nulls π_1 is not too large. This is expected: when the fraction of non-nulls is large, ordering is less relevant.

In particular, for small π_1, the gain in power can be as large as 20% (for $\sigma^2 = 1$) and as 30%

[6] The procedure of [LB16] is designed as to reject the first $\hat{k}$ null hypotheses, and accept the remaining $n - \hat{k}$. However, this specific structure is a design choice, and is not a constraint arising from the application.

[7] Note that ordering by increasing p'_j is equivalent to ordering by decreasing $|Z'_j|$ and the latter can be done without knowledge of the noise variance σ^2.

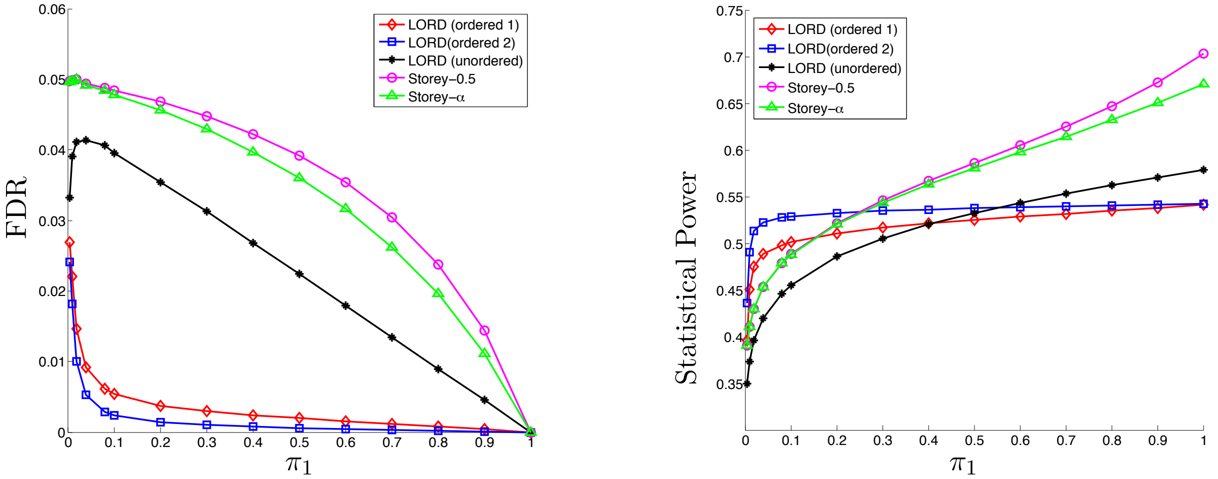


Figure 3: FDR and statistical power for LORD with favorably ordered hypotheses (setup of Section 5.2). Here $n = 3000$, $\pi_1 = 0.2$ and data are obtained by averaging over $20,000$ trials. Unordered: Null and non-null hypotheses are ordered at random. Ordered 1: hypotheses are ordered using very noisy side information ($\sigma^2 = 1$). Ordered 2: hypotheses are ordered using less noisy side information ($\sigma^2 = 1/2$).

(for $\sigma^2 = 1/2$). The resulting power is superior to adaptive BH [Sto02] for $\pi_1 \lesssim 0.15$ (for $\sigma^2 = 1$), or $\pi_1 \lesssim 0.25$ (for $\sigma^2 = 1/2$).

5.3 FDR control versus mFDR control

Aharoni and Rosset [AR14] proved that generalized alpha investing rules control mFDR_{w_0/b_0}. Formally,

$$\mathrm{mFDR}_{w_0/b_0}(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \frac{\mathbb{E}\, V^{\theta}(n)}{\mathbb{E}\, R(n) + (w_0/b_0)} \le b_0\,. \tag{43}$$

As mentioned before (see also Appendix A) this metric has been criticized because the two random variables $V^{\theta}(n)$ and $R(n)$ could be unrelated, while their expectations are close.

Our Theorem 3.3 controls a different metric that we called $\mathrm{sFDR}_{\eta}(n)$:

$$\mathrm{sFDR}_{w_0/b_0}(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\Big\{\frac{V^{\theta}(n)}{R(n) + (w_0/b_0)}\Big\} \le b_0\,. \tag{44}$$

This quantity is the expected ratio, and hence passes the above criticism. Note that both theorems yield control at level $\alpha = b_0$, for the same class of rules.

Finally, Theorem 3.1 controls a more universally accepted metric, namely FDR, at level $\alpha = w_0 + b_0$. A natural question is whether, in practice, we should choose w_0, b_0 as to guarantee FDR control (and hence set $w_0 + b_0 \le \alpha$) or instead be satisfied with mFDR and sFDR control, which allow for $b_0 = \alpha$ and hence potentially larger statistical power.

While an exhaustive answer to this question is beyond the scope of this paper, we repeated the simulations in Figure 1, using the two different criteria. The results (cf. Appendix A) suggest that this question might not have a simple answer. On one hand, under the setting of Figure 1 (independent p-values, large number of discovery) mFDR and sFDR seem stringent enough criteria. On the other, the gain in statistical power that is obtained from these criteria, rather than FDR, is somewhat marginal.

6 Control of False Discovery Exceedance

Ideally, we would like to control the proportion of false discoveries in any given realization of our testing procedures. We recall that this is given by (cf. Eq. (21))

$$\mathrm{FDP}^{\theta}(n) \equiv \frac{V^{\theta}(n)}{R(n) \vee 1}\,. \tag{45}$$

False discovery rate is the *expected* proportion of false discoveries. However –in general– control of FDR does not prevent FDP from varying , even when its average is bounded. In real applications, the actual FDP might be far from its expectation. For instance, as pointed out by [Bic04, Owe05], the variance of FDP can be large if the test statistics are correlated.

Motivated by this concern, the *false discovery exceedance* is defined as

$$\mathrm{FDX}_{\gamma}(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\big(\mathrm{FDP}^{\theta}(n) \ge \gamma\big)\,. \tag{46}$$

for a given tolerance parameter $\gamma \ge 0$. Controlling FDX instead of FDR gives a stronger preclusion from large fractions of false discoveries.

Several methods have been proposed to control FDX in an offline setting. Van der Laan, Dudoit and Pollard [vdLDP04] observed that any procedure that controls FWER, if augmented by a sufficiently small number of rejections, also controls FDX. Genovese and Wasserman [GW06] suggest controlling FDX by inverting a set of uniformity tests on the vector of p-values. Lehmann and Romano [LR12] proposed a step-down method to control FDX.

A natural criterion to impose in the online setting would be the control of $\sup_{n\ge 1}\mathrm{FDX}_{\gamma}(n)$. However, this does not preclude the possibility of large proportions of false discoveries at some (rare) random times n. It could be –as a cartoon example– that $\mathrm{FDP}^{\theta}(n) = 1/2$ independently with probability α at each n, and $\mathrm{FDP}^{\theta}(n) = \gamma/2$ with probability $1-\alpha$. In this case $\sup_{n\ge 1}\mathrm{FDX}_{\gamma}(n) \le \alpha$ but $\mathrm{FDP}^{\theta}(n) = 1/2$ almost surely for infinitely many times n. This is an undesirable situation.

A more faithful generalization of FDX to the online setting is therefore

$$\mathrm{FDX}_{\gamma} \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{P}\big(\sup_{n\ge 1} \mathrm{FDP}^{\theta}(n) \ge \gamma\big)\,. \tag{47}$$

We will next propose a class of generalized alpha-investing rules for online control of FDX_{γ}.

6.1 The effect of reducing test levels

Before describing our approach, we demonstrate through an example that the FDP can differ substantially from its expectation. We also want to illustrate how a naive modification of the previous rules only achieves a better control of this variability at the price of a significant loss in power.

Note that the desired bound $\mathrm{FDP}^{\theta}(n) < \gamma$ follows if we can establish $b_0 R(n) - V(n) + (\gamma - b_0) > 0$ for some $\gamma \ge b_0 \ge 0$. Recall that a generalized alpha-investing procedure continues until the potential $W(n)$ remains non-negative. Therefore, for such a procedure, it suffices to bound the probability that the stochastic process $B(n) \equiv b_0 R(n) - W(n) - V(n) + (\gamma - b_0)$ crosses zero. As we show in Lemma E.1, $B(n)$ is a submartingale, and thus in expectation it moves away from zero.

In order to bound the deviations from the expectation, consider the submartingale increments $B_j \equiv B(j) - B(j-1)$ given by

$$B_j = (b_0 - \psi_j)R_j + \varphi_j - V_j \,. \tag{48}$$

If the j-th null hypothesis is false, i.e. $\theta_j \neq 0$, we have $V_j = 0$ and $B_j \geq 0$ invoking assumption G1 and noting that $R_j \in \{0, 1\}$. Under the null hypothesis, $V_j = R_j$, and

$$\mathrm{Var}(B_j|\mathcal{F}_{j-1}) = (b_0 - \psi_j - 1)^2 \alpha_j (1 - \alpha_j) \,. \tag{49}$$

Reducing $\mathrm{Var}(B_j|\mathcal{F}_{j-1})$ lowers variations of the submartingale and hence the variation of the false discovery proportions. Note that for a generalized alpha-investing rule, if we keep b_0, ψ_j unchanged and lower the test levels α_j, the rule still satisfies conditions G1 , G2 and thus controls FDR at the desired level. On the other hand, this modification decreases $\mathrm{Var}(B_j|\mathcal{F}_{j-1})$ as per Eq (49).

In summary, reducing the test levels has the effect of reducing the variation of false discovery proportion at the expense of reducing statistical power.

We carry out a numerical experiment within a similar setup as the one discussed in Section 5. A set of n hypotheses are tested, each specifying mean of a normal distribution, $H_j : \theta_j = 0$. The test statistics are independent, normally distributed random variables $Z_j \sim \mathsf{N}(\theta_j, 1)$. For non-null hypotheses, we set $\theta_j = 3$. The total number of tests is $n = 1000$ of which the first 100 are non-null.

We consider three different testing rules, namely alpha investing, alpha spending with rewards and LORD , all ensuring FDR control at level $\alpha = 0.05$. The details of these rules as well as the choice of parameters is the same as Section 5.

In order to study the effect of reducing test levels, for each of these rules we truncate them by a threshold value T, i.e. we use $\alpha_j^T = \alpha_j \vee T$. We plot the histogram of false discovery proportions using $30,000$ replications of the test statistics sequence. We further report standard deviation and 0.95 quantile of FDPs. The results are shown in Figs. 4, 5, 6.

As a first remark, while all of the rules considered control FDR below $\alpha = 0.05$, the actual false discovery proportion in Figs. 4, 5, 6 has a very broad distribution. Consider for instance alpha-investing, at threshold level $T = 0.9$. Then FDP exceeds 0.15 (three times the nominal value) with probability 0.13.

Next we notice that reducing the test levels (by reducing T) has the desired effect of reducing the variance of the FDP. This effect is more pronounced for alpha-investing. Nevertheless quantifying this effect is challenging due to the complex dependence between B_j and history $\mathcal{F}_{j-1}$. This makes it highly nontrivial to adjust threshold T to obtain $\mathrm{FDX}_\gamma \leq \alpha$. In the next section we achieve this through a different approach.

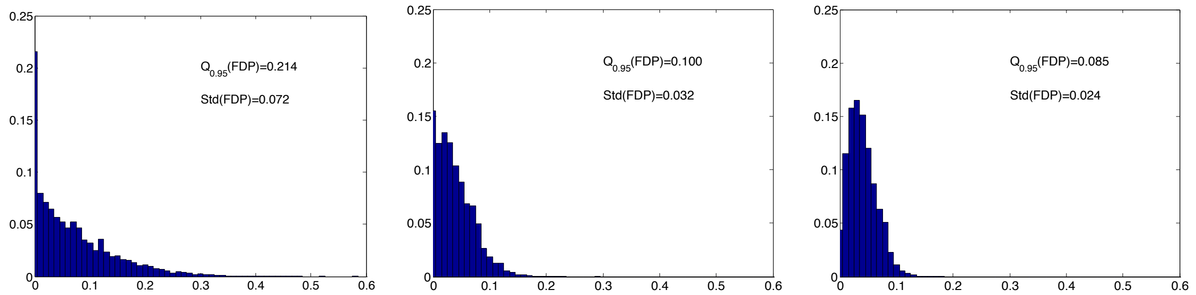
6.2 Rules for controlling FDX_γ

Let $M(0) = \gamma - b_0 - w_0 > 0$ and define, for $n \in \mathbb{N}$, $M(n) = M(0) + \sum_{j=1}^n M_j$, where

$$M_j \equiv \max\{(1 + \psi_j - b_0)(\alpha_j - R_j), (b_0 - \psi_j)R_j, \psi_j - b_0\} \,. \tag{50}$$

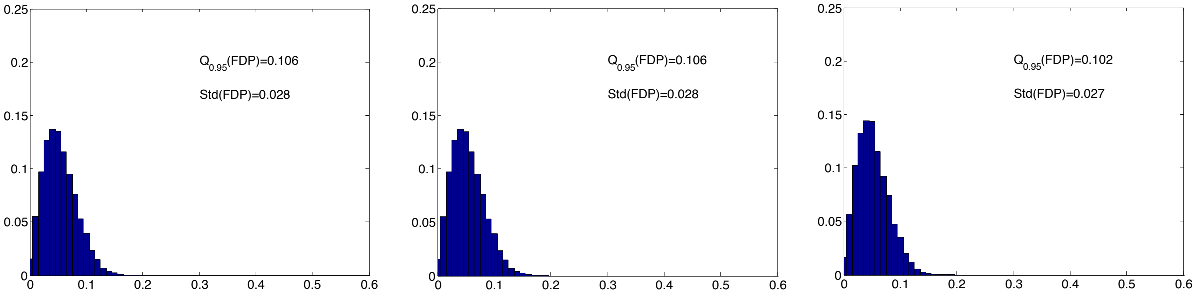
Note that $M(n)$ is a function of $(R_1, \dots, R_n)$, i.e. it is measurable on $\mathcal{F}_n$. We then require the following conditions in addition to G1 and G2 introduced in Section 2.1:

G3. $w_0 < \gamma - b_0$.



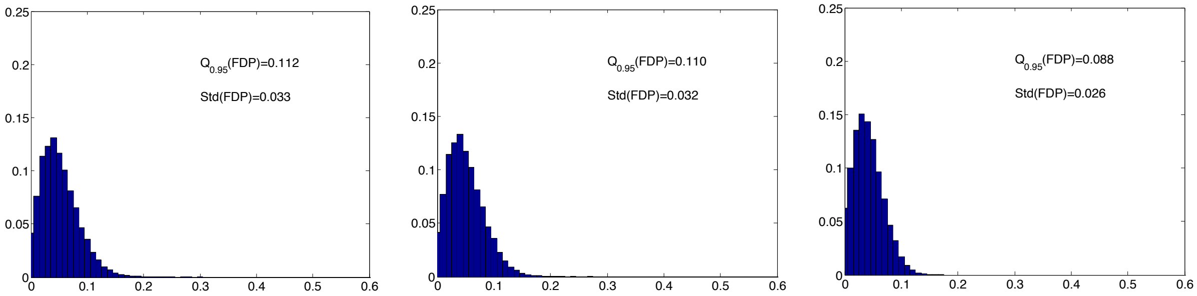
(a) $T = 0.9$ (b) $T = 0.6$ (c) $T = 0.3$

Figure 4: Histogram of FDP for alpha investing rule with different values of T



(a) $T = 0.9$ (b) $T = 0.6$ (c) $T = 0.3$

Figure 5: Histogram of FDP for alpha spending with rewards for different values of T



(a) $T = 0.9$ (b) $T = 0.6$ (c) $T = 0.3$

Figure 6: Histogram of FDP for LORD rule with different values of T

G4. For $j \in \mathbb{N}$ and all $\boldsymbol{R}_1^j \in \{0,1\}^j$, if

$$M(j) + \xi_{j+1} > \frac{\gamma - b_0 - w_0}{1-\alpha}\,, \tag{51}$$

then $\alpha_i = 0$ for all $i > j$, where we define $\xi_j \equiv \max\{(1+\psi_j - b_0)\alpha_j, |b_0 - \psi_j|\}$.

(This condition is well posed since $M(j)$ and ξ_{j+1} are functions of $\boldsymbol{R}_1^j$.)

Note that any generalized alpha-investing rule can be modified as to satisfy these conditions. Specifically, the rule keeps track of LHS of (51) (it is an *observable* quantity) and whenever inequality (51) is violated, the test levels are set to zero onwards, i.e, $\alpha_i = 0$ for $i \ge j$. The sequence $(\xi_j)_{j\in\mathbb{N}}$ is constructed in a way to be a predictable process that bounds M_j. Consequently, $M(j) + \xi_{j+1} \in \mathcal{F}_j$ bounds $M(j+1)$.

The decrement and increment values φ_j and ψ_j are determined in way to satisfy conditions G2 and G5.

We then establish FDX control under a certain negative dependency condition on the test statistics.

Theorem 6.1. *Assume that the p-values $(p_i)_{i\in\mathbb{N}}$ are such that, for each $j \in \mathbb{N}$, and all $\boldsymbol{\theta} \in H_j$ (i.e. all $\boldsymbol{\theta}$ such that the null hypothesis $\theta_j = 0$ holds), we have*

$$\mathbb{P}_{\boldsymbol{\theta}}(p_j \le \alpha_j | \mathcal{F}_{j-1}) \le \alpha_j\,, \tag{52}$$

almost surely.

Then, any generalized alpha-investing rule that satisfies conditions G3 *,* G4 *above (together with* G1 *and* G2 *) controls the false discovery exceedance:*

$$\text{FDX}_\gamma \le \alpha\,. \tag{53}$$

The proof of this theorem is presented in Appendix E. Notice that the dependency condition (52) is satisfied, in particular, if the p-values are independent.

Example 6.2. For given values of $\alpha \in (0,1)$ and $\gamma \in (\alpha, 1)$, consider LORD algorithm with $b_0 = \alpha$, $\psi_j = \alpha$ for $j \in \mathbb{N}$ and $w_0 = (\gamma - \alpha)/2$. By Eq. (50), we have $M_j = \alpha_j \mathbb{I}(R_j = 0)$. In order to satisfy condition G4 , the rule keeps track of $M(n)$ and stops as soon as inequality (51) is violated, i.e.,

$$\alpha_{n+1} + \sum_{i=1}^{n} \alpha_i \mathbb{I}(R_i = 0) > \frac{\gamma - \alpha}{2(1-\alpha)}\,. \tag{54}$$

Note that for LORD , the potential sequence $W(n)$ always remain positive and thus the stopping criterion is defined solely based on the above inequality. Clearly, this rule satisfies assumptions G1 , G2 , G3 , G4 and by applying Theorem 6.1 ensures $\text{FDX}_\gamma \le \alpha$.

We use the above rule to control false discovery exceedance for the simulation setup described in Section 6.1 for values of $\alpha = 0.05$ and $\gamma = 0.15$. The results are summarized in Table 1. The false discovery rates and proportions are estimated using $30,000$ realizations of test statistics. As we see the rule controls both FDR and FDX_γ below α.

Online control of FDX$_\gamma$ using stopping criterion (54)					
π_1	0.005	0.01	0.02	0.03	0.04
FDX$_\gamma$	0.028	0.004	0.000	0.000	0.000
FDR	0.006	0.005	0.005	0.005	0.005
Power	0.666	0.699	0.679	0.658	0.639

Table 1: FDX$_\gamma$ and FDR for Lord with stopping criterion (54) using 30,000 realizations of the test statistics. Here, $\alpha = 0.05$ and $\gamma = 0.15$, and π_1 represents the fraction of truly non-null hypotheses that appear at the beginning of the stream as described in Section 6.1.

7 Diabetes prediction data

In order to explore a more realistic setting, we apply online testing to a health screening example. The adoption of electronic health records has been accelerating in recent years both because of regulatory and technological forces. A broadly anticipated use of these data is to compute predictive health scores [BSOM$^+$14]. A high-risk score for some chronic disease can trigger an intervention, such as an incentive for healthy behavior, additional tests, medical follow-up and so on. Predictive scores for long term health status are already computed within wellness programs that are supported by many large employers in the US.

To be concrete, we will consider a test to identify patients that are at risk of developing diabetes (see also [BBM15] for related work). Notice that a positive test outcome triggers an intervention (e.g. medical follow-up) that cannot be revised, and it is important to control the fraction of alerts that are false discoveries. It is therefore natural to view this as an online hypothesis testing problem as the ones considered in the previous sections. For each patient j, we form the null hypothesis H_j: “The patient will not develop diabetes” versus its alternative.

We use a diabetes dataset released by Practice Fusion as part of a Kaggle competition[8]. We only use the ‘train’ portion of this dataset, which contains de-identified medical records of $n_{\text{tot}} = 9,948$ patients. For each of patient, we have a response variable that indicates if the patient is diagnosed with Type 2 diabetes mellitus, along with information on medications, lab results, immunizations, allergies and vital signs.

We develop a predictive score based on the available records, and will generalized alpha-investing rules to control FDR in these tests.

We split the data in three sets Train1, comprising 60% of the data, Train2, 20% of the data, and Test, 20% of the data. The Train sets are used to construct a model, which allows to compute p-values. The p-values are then used in an online testing procedure applied to the Test set. In detail, we proceed as follows:

Feature extraction. For each patient i, we denote by $y_i \in \{0, 1\}$ the response variable (with $y_i = 1$ corresponding to diagnosis of diabetes) and we construct a vector of covariates $\boldsymbol{x}_i \in \mathbb{R}^d$, $d = 805$ by using the following attributes:

- *Transcript records*: Year of birth, gender, and BMI
- *Diagnosis information*: We include 80 binary features corresponding to different ICD-9 codes.

[8] See http://www.kaggle.com/c/pf2012-diabetes.

- *Medications*: We include 80 binary features indicating the use of various medications.
- *Lab results*: For 70 lab test observations we include a binary feature indicating whether the patient has taken the test. We further includes abnormality flags and the observed outcomes as features. In addition, we bin the outcomes into 10 quantiles and make 10 binary features via one-hot encoding.

Construction of logistic model. We use a logistic link function to model the probability of developing diabetes. Let us emphasize that we do not did not optimize the link function. The primary goal of this section is to show applicability of *online* multiple testing setup in many real world problems.

Explicitly, we model the probability of no diabetes as

$$\mathbb{P}(Y_i = 0|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{1}{1 + e^{\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle}} \,. \tag{55}$$

The parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is estimated using the $\mathsf{Train1}$ data set.

Construction of p-values. Let T_0 be the subset of $\mathsf{Train2}$ with $Y_i = 0$, and let $n_0 \equiv |T_0|$. For $i \in T_0$, we compute the predictive score $q_i = 1/(1 + e^{\langle \boldsymbol{\theta}, \boldsymbol{x}_i \rangle})$. For each $j \in \mathsf{Test}$, we compute $q_j^{\mathsf{Test}} = 1/(1 + e^{\langle \boldsymbol{\theta}, \boldsymbol{x}_j \rangle})$, and construct a p-value p_j by

$$p_j \equiv \frac{1}{n_0} \big|\big\{\, i \in T_0 : \; q_i \le q_j^{\mathsf{Test}} \big\}\big| \,. \tag{56}$$

For patients with high-risk of a diabetes (according to the model), q_j is small resulting in small p-values, as expected. We will use these p-values as input to our online testing procedure.

Note that, since the $\mathsf{Train1}$ and Test sets are exchangeable, the null p-values will be uniform in expectation (and asymptotically uniform under mild conditions).

Online hypothesis testing. We consider several online hypothesis testing procedures aimed at controlling FDR below a nominal value $\alpha = 0.1$ (in particular, without adjusting for dependency among the p-values). For LORD , we choose the sequence $\boldsymbol{\gamma} = (\gamma_m)_{m \ge 1}$ following Eq. (41). For each rule, we compute corresponding false discovery proportion (FDP) and statistical power and average them over 20 random splittings of data into $\mathsf{Train1}$, $\mathsf{Train2}$, Test. Table 2 summarizes the results. As we see, generalized alpha-investing rules control FDR close to the desired level, and have statistical power comparable to the benchmark offline approaches.

Result for Diabetes data

	FDR	Power
LORD	0.126	0.531
Alpha-investing	0.141	0.403
Bonferroni	0.195	0.166
Benjamin-Hochberg (BH)	0.121	0.548
Adaptive BH (Storey-α)	0.123	0.569
Adaptive BH (Storey-0.5)	0.128	0.573

Table 2: False discovery rate (FDR) and statistical power for different online hypotheses testing rules for the diabetes dataset. The reported numbers are averages over 20 random splits of data into training set and test set.

8 Discussion and further related work

We list below a few lines of research that are related to our work.

General context. An increasing effort was devoted to reducing the risk of fallacious research findings. Some of the prevalent issues such as publication bias, lack of replicability and multiple comparisons on a dataset were discussed in Ioannidis's 2005 papers [Ioa05b, Ioa05a] and in [PSA11].

Statistical databases. Concerned with the above issues and the importance of data sharing in the genetics community, [RAN14] proposed an approach to public database management, called Quality Preserving Database (QPD). A QPD makes a shared data resource amenable to perpetual use for hypothesis testing while controlling FWER and maintaining statistical power of the tests. In this scheme, for testing a new hypothesis, the investigator should pay a price in form of additional samples that should be added to the database. The number of required samples for each test depends on the required effect size and the power for the corresponding test. A key feature of QPD is that type I errors are controlled at the management layer and the investigator is not concerned with p-values for the tests. Instead, investigators provide effect size, assumptions on the distribution of the data, and the desired statistical power. A critical limitation of QPD is that all samples, including those currently in the database and those that will be added, are assumed to have the same quality and are coming from a common underlying distribution. Motivated by similar concerns in practical data analysis, [DFH$^+$14] applies insights from differential privacy to efficiently use samples to answer adaptively chosen estimation queries. These papers however do not address the problem of controlling FDR in online multiple testing.

Online feature selection. Building upon alpha-investing procedures, [LFU11] develops VIF, a method for feature selection in large regression problems. VIF is accurate and computationally very efficient; it uses a one-pass search over the pool of features and applies alpha-investing to test each feature for adding to the model. VIF regression avoids overfitting due to the property that alpha-investing controls mFDR. Similarly, one can incorporate Lord in VIF regression to perform fast online feature selection and provably avoid overfitting.

High-dimensional and sparse regression. There has been significant interest over the last two years in developing hypothesis testing procedures for high-dimensional regression, especially in conjunction with sparsity-seeking methods. Procedures for computing p-values of low-dimensional coordinates were developed in [ZZ14, VdGBRD14, JM14a, JM14b, JM13]. Sequential and selective inference methods were proposed in [LTTT14, FST14, TLTT14]. Methods to control FDR were put forward in [BC15, BvdBS$^+$15].

As exemplified by VIF regression, online hypothesis testing methods can be useful in this context as they allow to select a subset of regressors through a one-pass procedure. Also they can be used in conjunction with the methods of [LTTT14], where a sequence of hypothesis is generated by including an increasing number of regressors (e.g. sweeping values of the regularization parameter).

In particular, [GWCT15, LB16] develop multiple hypothesis testing procedures for ordered tests. Note, however, that these approaches fall short of addressing the issues we consider, for several reasons: (i) They are not online, since they reject the first $\hat{k}$ null hypotheses, where $\hat{k}$ depends on all the p-values. (ii) They require knowledge of all past p-values (not only discovery events) to compute the current score. (iii) Since they are constrained to reject all hypotheses before $\hat{k}$, and accept them after, they cannot achieve any discovery rate increasing with n, let alone nearly linear in n. For instance in the mixture model of Section 4, if the fraction of true non-null is $\pi_1 < \alpha$, then the methods of [GWCT15, LB16] achieves $O(1)$ discoveries out of $\Theta(n)$ true non-null. In other words their power is of order $1/n$ in this simple case.

Acknowledgements

A.M. was partially supported by NSF grants CCF-1319979 and DMS-1106627 and the AFOSR grant FA9550-13-1-0036.

A FDR **versus** mFDR

The two main criteria discussed in the present paper are $\mathrm{FDR}(n)$ and $\mathrm{mFDR}_\eta(n)$ at level $\eta = w_0/b_0$. Recall that these are formally defined by

$$\mathrm{FDR}(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\Big\{\frac{V^\theta(n)}{R(n)\vee 1}\Big\}\,, \tag{57}$$

$$\mathrm{mFDR}_\eta(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \frac{\mathbb{E}\,V^\theta(n)}{\mathbb{E}\,R(n)+\eta}\,. \tag{58}$$

In addition, we introduced a new metric, that we called $\mathrm{sFDR}_\eta(n)$ (for smoothed FDR):

$$\mathrm{sFDR}_\eta(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\Big\{\frac{V^\theta(n)}{R(n)+\eta}\Big\}\,. \tag{59}$$

Note that mFDR is different from the other criteria in that it does not control the probability of a property of the realized set of tests; rather it controls the ratio of expected number of false discoveries to the expected number of discoveries. In this appendix we want to document two points already mentioned in the main text:

1. FDR and mFDR can be –in general– very different. More precisely, we show through a numerical simulation that controlling mFDR does not ensure controlling FDR at a similar level. This provides further motivation for Theorem 3.1.

   We discuss this point in Section A.1.

2. Theorem 3.1 establishes $\mathrm{FDR}(n) \le b_0 + w_0$ and Theorem 3.3 ensures $\mathrm{sFDR}_{w_0/b_0}(n) \le b_0$. Analogously, [AR14] proved $\mathrm{mFDR}_{w_0/b_0}(n) \le b_0$. In other words, if we target mFDR or sFDR control, we can use larger values of w_0 and hence –potentially– achieve larger power.

   We explore this point in Section A.2.

A.1 FDR **and** mFDR **can be very different**

Example A.1. Since Theorem 3.1 shows that generalized alpha-investing procedures *do control* FDR, our first example will be of different type. Indeed, since we want to show that *in general* FDR and mFDR are very different, we will consider a very simple rule.

We observe $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ where $X_j = \theta_j + \varepsilon_j$ and we want to test null hypotheses $H_j : \theta_j = 0$. The total number of tests is $n = 3,000$ from which the first $n_0 = 2,700$ hypotheses are null and the remaining are non-null. For null cases, $X_1, X_2, \dots, X_{n_0}$ are independent $\mathsf{N}(0,1)$ observations. Under the alternative, we assume $\theta_j = 2$ and $(\varepsilon_{n_0+1}, \dots, \varepsilon_n)$ follows a multivariate normal distribution with covariance $\Sigma = \rho\mathbf{1}\mathbf{1}^\mathsf{T} + (1-\rho)\mathrm{I}$, with $\mathbf{1}$ the all-one vector. Here ρ controls the dependency among the non-null test statistics. In our simulation, we set $\rho = 0.9$. It is worth noting that this setting is relevant to many applications as it is commonly observed that the non-null cases are clustered.

We consider a single step testing procedure, namely

$$R_j = \begin{cases} 1 & \text{if } |X_i| \le t\,, \\ 0 & \text{if } |X_i| > t\,. \end{cases} \tag{60}$$

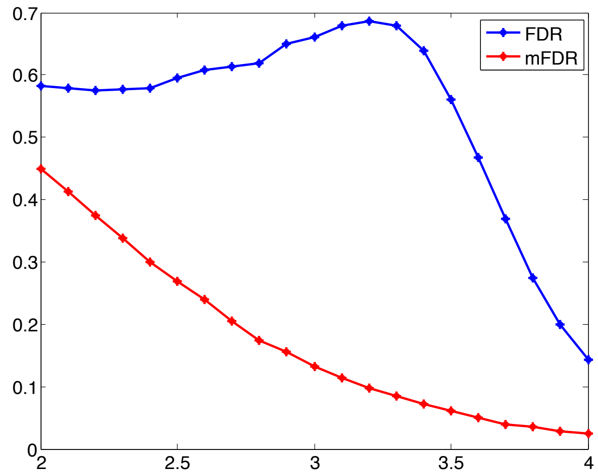


Figure 7: FDR and mFDR for the single step procedure of Eq. (60) under the setting of Example A.1.

The value of t is varied from 2 to 4 and mFDR and FDR are computed by averaging over 10^4 replications. The result is shown in Fig. 7. As we see the two measures are very different. For instance, choosing $t = 3$ controls mFDR below $\alpha = 0.2$, but results in FDR $\gtrsim 0.6$.

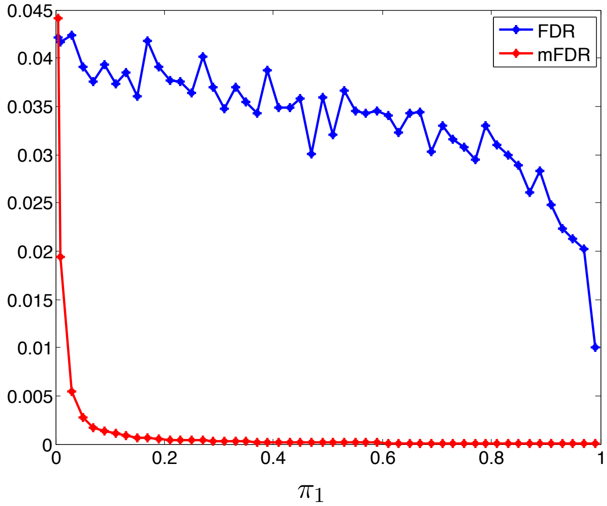


Figure 8: FDR and mFDR for alpha investing rule under the setting of Example A.1 for various fraction of non-null hypotheses π_1.

Example A.2. We next consider the alpha investing rule, as described in Subsection 2.2.1 with α_j set based on equation (40), at nominal value $\alpha = 0.05$. In this case Theorem 3.1 guarantees FDR $\le \alpha$. However FDR and mFDR can still be very different as demonstrated in Figure 8.

The hypothesis testing problem is similar to the one in the previous example. We consider a normal vector $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$, $X_i = \theta_i + \varepsilon_i$, $n = 3,000$, and want to test for the null hypotheses $\theta_i = 0$. The noise covariance has the same structure as in the previous example, and the means are $\theta_j = 4$ when the null is false. Unlike in the previous example, we consider a varying proportion π_1 of non-zero means. Namely, the null is false for $i \in \{n_0+1,\dots,n\}$, with $(n-n_0)=\pi_1 n$.

The results in Fig. [FS07] are obtained by averaging over 10^4 replications. Alpha investing controls $\mathrm{mFDR}, \mathrm{FDR} \le 0.05$, as expected (Indeed conditions of Theorem 3.1 hold in this example since the p-values of true nulls are independent from other p-values). However, the two metrics are drastically different and a bound on mFDR does not imply a bound on FDR at the same level. For instance, at $\pi_1 = 0.1$ we have $\mathrm{mFDR} \lesssim 0.001$ while $\mathrm{FDR} \approx 0.04$.

A.2 Comparing FDR and mFDR with respect to statistical power

In Figure 9 we simulated two generalized alpha investing rules, namely Lord and simple alpha investing [FS07], under the same setting of Section 5.1, with Gaussian alternatives, and compare two different choices of the parameters w_0 (initial wealth) and b_0 (bound on the reward function in Eqs. (9), (10)):

> *Solid lines.* Correspond to the choice already used in Section 5.1, namely $w_0 = 0.005$, $b_0 = 0.045$. By Theorem 3.1, this is guaranteed to control $\mathrm{FDR} \le 0.05$.
>
> *Dashed lines.* Correspond to a more liberal choice, $w_0 = 0.05$, $b_0 = 0.05$. By [AR14, Theorem 1], this controls $\mathrm{mFDR}_1 \le 0.05$. Theorem 3.3 provides the additional guarantee

$$\mathrm{sFDR}_1(n) \equiv \sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}\Big\{\frac{V^\theta(n)}{R(n)+1}\Big\} \le 0.05\,. \tag{61}$$

In Figure 9 we compare FDR, sFDR and statistical power for these two choice. As expected FDR and sFDR are slightly higher for the second choice, but still FDR appears to be below the target value $\alpha = 0.05$. This is to be expected on the basis of Remark 3.2 which implies $\mathrm{FDR} \lesssim b_0$ when the number of discoveries $R(n)$ is large with high probability. We conclude that –in the case of non-nulls arriving at random, and sufficiently many strong signals– mFDR control is conservative enough.

The last panel in the same figure shows the increase in power obtained by the second choice of parameters that targets mFDR control. Note that the advantage is –in this example– somewhat marginal. In other words, FDR control can be guaranteed without incurring large losses in power.

B FDR for independent p-values: Proof of Theorem 3.1 and Theorem 3.3

Lemma B.1. *Assume the p-values $p_1,\dots p_n$ to be independent, and that $\theta_j = 0$ (i.e. p_j is a true null p-value). Let $R(n) = \sum_{i=1}^n R_i$ be the total number of rejection up until time n for a monotone online rule. Let $f : \mathbb{Z}_{\ge 0} \to \mathbb{R}_{\ge 0}$ be a non-increasing, non-negative function on the integers. Then*

$$\mathbb{E}\Big\{\mathbb{I}\{p_j \le \alpha_j\}\, f(R(n))\Big|\mathcal{F}_{j-1}\Big\} \le \mathbb{E}\Big\{\alpha_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big\}\,. \tag{62}$$

Proof. We let $\boldsymbol{p} = (p_1,p_2,\dots,p_n)$ be the sequence of p-values until time n, and denote by $\widetilde{\boldsymbol{p}} = (p_1,p_2,\dots,p_{j-1},0,p_{j+1}\dots,p_n)$ the vector obtained from $\boldsymbol{p}$ by setting $p_j = 0$. We let $\boldsymbol{R} = (R_1,R_2,\dots,R_n)$ be the sequence of decisions on input $\boldsymbol{p}$, and denote by $\widetilde{\boldsymbol{R}} = (\widetilde{R}_1,\widetilde{R}_2,\dots,\widetilde{R}_n)$

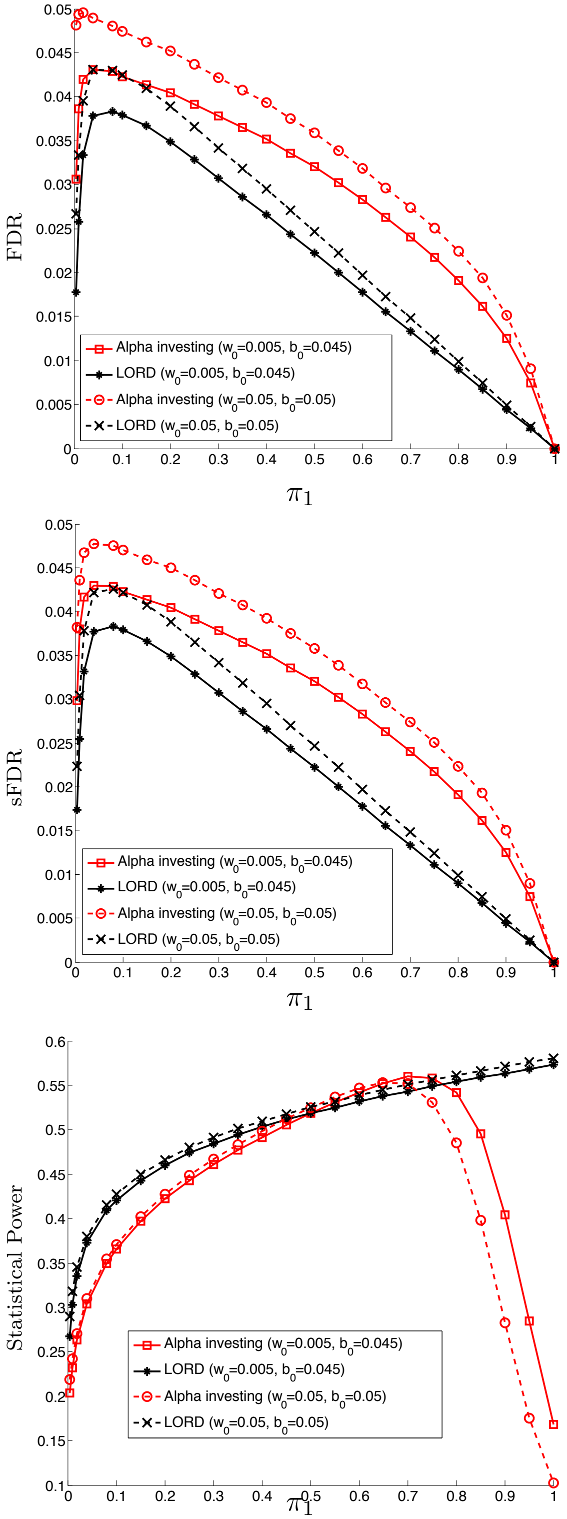


Figure 9: FDR (top), sFDR (center), and statistical power (bottom) versus fraction of non-null hypotheses π_1, for the Gaussian setup described in Section 5. Solid lines: parameters are tuned to control FDR. Dashed lines: parameters are tuned to control mFDR and sFDR.

the sequence of decision when the same rule is applied to input $\widetilde{\boldsymbol{p}}$. The total numbers of rejections are denoted by $R(n)$ and $\widetilde{R}(n)$. Finally, let α_ℓ and $\widetilde{\alpha}_\ell$ denote test levels given by the rule applied to $\boldsymbol{p}$ and $\widetilde{\boldsymbol{p}}$, respectively.

Since $\alpha_\ell = \alpha_\ell(\boldsymbol{R}_1^{\ell-1})$, $\widetilde{\alpha}_\ell = \alpha_\ell(\widetilde{\boldsymbol{R}}_1^{\ell-1})$, we have $\alpha_\ell = \widetilde{\alpha}_\ell$ for $\ell \le j$. Observe that on the event $\{p_j \le \alpha_j\}$, we have $R_j = \widetilde{R}_j$ and therefore $\alpha_\ell = \widetilde{\alpha}_\ell$ for all $1 \le \ell \le n$. In words, when H_j is rejected, the actual value of p_j does not matter. Therefore, on the same event, we have $R(n) = \widetilde{R}(n)$, whence

$$\mathbb{I}\{p_j \le \alpha_j\}\, f(R(n)) = \mathbb{I}\{p_j \le \alpha_j\}\, f(\widetilde{R}(n))\,. \tag{63}$$

Taking conditional expectations

$$\begin{aligned}\mathbb{E}\Big\{\mathbb{I}\{p_j \le \alpha_j\}\, f(R(n))\Big|\mathcal{F}_{j-1}\Big\} &= \mathbb{E}\Big\{\mathbb{I}\{p_j \le \alpha_j\}\, f(\widetilde{R}(n))\Big|\mathcal{F}_{j-1}\Big\} &(64)\\ &= \mathbb{E}\Big\{\alpha_j\, f(\widetilde{R}(n))\Big|\mathcal{F}_{j-1}\Big\}\,, &(65)\end{aligned}$$

where we used the fact that, conditional on $\mathcal{F}_{j-1} = \sigma(R_1, \dots, R_{j-1})$, level α_j is deterministic (it is measurable on $\mathcal{F}_{j-1}$). Further, p_j is independent of the other p-values and thus in particular is independent of the sigma-algebra generated by $\mathcal{F}_{j-1} \cup \sigma(\widetilde{R}(n))$.

Note that $\widetilde{R}_j = 1$ and by monotonicity of the rule, $\widetilde{\alpha}_{j+1} \ge \alpha_{j+1}$ and hence $\widetilde{R}_{j+1} \ge R_{j+1}$. Repeating this argument, we obtain $\widetilde{\boldsymbol{R}} \succeq \boldsymbol{R}$ which implies $\widetilde{R}(n) \ge R(n)$. Hence, equation (65) yields the desired result. □

We next prove Theorem 3.1 and Theorem 3.1. The argument can be present in a unified way, applying Lemma B.1 to two different choices of the function $f(\,\cdot\,)$.

Proof (Theorem 3.1 and Theorem 3.1). As above, $R(j) = \sum_{i=1}^{j} R_i$ denotes the number of discoveries up until time j, and $V(j)$ the number of false discoveries among them. Fixing $f : \mathbb{Z}_{\ge 0} \to \mathbb{R}_{\ge 0}$ a non-negative, non-increasing function, we define the sequence of random variables

$$A(j) \equiv \big\{b_0 R(j) - V(j) - W(j)\big\} f(R(n))\,, \tag{66}$$

indexed by $j \in \{1, 2, \dots, n\}$. First of all, note that $A(j)$ is integrable. Indeed $0 \le R(j), V(j) \le j$, and $W(j)$ takes at most 2^j finite values (because it is a function of $\boldsymbol{R}_1^j \in \{0,1\}^j$).

Let $A_j = A(j) - A(j-1)$, $V_j = V(j) - V(j-1)$ and $W_j = W(j) - W(j-1) = -\varphi_j + R_j\psi_j$. Hence, $A_j = \{(b_0 - \psi_j)R_j - V_j + \varphi_j\}\, f(R(n))$. Assuming $\theta_j = 0$ (i.e. the j-th null hypothesis H_j is true) we have $R_j = V_j$ and thus $A_j = \{(b_0 - \psi_j - 1)R_j + \varphi_j\}\, f(R(n))$. Taking conditional expectation of this quantity (and recalling that φ_j, ψ_j are measurable on $\mathcal{F}_{j-1}$), we get

$$\begin{aligned}\mathbb{E}(A_j|\mathcal{F}_{j-1}) &= (b_0 - \psi_j - 1)\mathbb{E}\Big(R_j f(R(n))\Big|\mathcal{F}_{j-1}\Big) + \mathbb{E}\Big(\varphi_j f(R(n))\Big|\mathcal{F}_{j-1}\Big) &(67)\\ &\ge (b_0 - \psi_j - 1)\mathbb{E}\Big(\alpha_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big) + \mathbb{E}\Big(\varphi_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big) &(68)\\ &= \mathbb{E}\Big\{\big((b_0 - \psi_j - 1)\alpha_j + \varphi_j\big)\, f(R(n))\Big|\mathcal{F}_{j-1}\Big\} &(69)\\ &\ge \mathbb{E}\Big\{\big((b_0 + 1 - b_0 - \varphi_j/\alpha_j - 1)\alpha_j + \varphi_j\big)\, f(R(n))\Big|\mathcal{F}_{j-1}\Big\} = 0\,. &(70)\end{aligned}$$

The first inequality holds because of Lemma B.1 and noting that $\psi_j \ge 0$ and $b_0 \le 1$. The last step follows from condition (10) that holds for generalized alpha-investing rules.

Assume next $\theta_j \neq 0$ (i.e. the j-th null hypothesis H_j is true). In this case $V_j = 0$, and therefore $A_j = \{(b_0 - \psi_j)R_j + \varphi_j\}\, f(R(n))$. Taking conditional expectation, we get

$$\begin{aligned}\mathbb{E}(A_j|\mathcal{F}_{j-1}) &= (b_0 - \psi_j)\mathbb{E}\Big(R_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big) + \mathbb{E}\Big(\varphi_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big) &(71)\\ &\geq (b_0 - \varphi_j - b_0)\mathbb{E}\Big(R_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big) + \mathbb{E}\Big(\varphi_j\, f(R(n))\Big|\mathcal{F}_{j-1}\Big) &(72)\\ &\geq \mathbb{E}\Big\{\big(-\varphi_j R_j + \varphi_j\big)\, f(R(n))\Big|\mathcal{F}_{j-1}\Big\} \geq 0\,, &(73)\end{aligned}$$

where the first inequality follows from condition (9) and in the last step we used the fact $R_j \leq 1$.

We therefore proved that $\mathbb{E}\{A_j|\mathcal{F}_{j-1}\} \geq 0$ irrespectively of θ_j. Since $V(0) = R(0) = 0$, we get $A(0) = -W(0)\, f(R(n)) \geq -w_0\, f(R(n))$. Therefore

$$\begin{aligned}\mathbb{E}\{A(n)\} &= \mathbb{E}\{A(0)\} + \sum_{j=1}^{n} \mathbb{E}\{A_j\} &(74)\\ &= -w_0\mathbb{E}\Big\{f(R(n))\Big\} + \sum_{j=1}^{n} \mathbb{E}\{\mathbb{E}(A_j|\mathcal{F}_{j-1})\} \geq -w_0\mathbb{E}\Big\{f(R(n))\Big\}\,. &(75)\end{aligned}$$

Using the definition of $A(n)$, and $R(n)/(R(n) \vee 1) \leq 1$, this implies

$$b_0\mathbb{E}\Big\{R(n)f(R(n))\Big\} - \mathbb{E}\Big\{V(n)\, f(R(n))\Big\}\mathbb{E}\Big\{W(n)\, f(R(n))\Big\} \geq -w_0\mathbb{E}\Big\{f(R(n))\Big\} \qquad (76)$$

Since $W(n) \geq 0$ by definition, this yields

$$\mathbb{E}\Big\{V(n)\, f(R(n))\Big\} \leq b_0\mathbb{E}\Big\{R(n)\, f(R(n))\Big\} + w_0\mathbb{E}\Big\{f(R(n))\Big\}\,. \qquad (77)$$

Substituting $f(R) = 1/(R \vee 1)$ we obtain the claim of Theorem 3.1 (as well as Remark 3.2).

By using instead $f(R) = 1/\{R + (w_0/b_0)\}$, we obtain Theorem 3.3. □

C A lower bound on FDR

In this section we prove Remark 3.4, stating that Theorems 3.1 and 3.3 cannot be substantially improved, unless we restrict to a subclass of generalized alpha-investing rules. In particular, Theorem 3.3 is optimal and Theorem 3.1 is sub-optimal at most by an additive term w_0. A formal statement is given below.

Proposition C.1. *For any $w_0, b_0 \geq 0$, there exist a generalized alpha-investing rule, with parameters w_0, b_0, and a sequence of p-values satisfying the assumptions of Theorem 3.1 such that*

$$\begin{aligned}&\liminf_{n\to\infty} \mathrm{FDR}(n) \geq b_0\,, &(78)\\ &\liminf_{n\to\infty} \mathbb{E}\Big\{\frac{V^\theta(n)}{R(n) + (w_0/b_0)}\Big\} \geq b_0\,. &(79)\end{aligned}$$

Proof. For the generalized alpha-investing rule, we use LORD with sequence of parameters $(\gamma_m)_{m\geq 1}$. We assume $\gamma_m > 0$ for all $m \geq 1$. Fix $m_0 \geq 2$. We construct the p-values $(p_i)_{i\geq 1}$ by assuming that H_j is false if $j \in \{m_0, 2m_0, 3m_0, \dots\} \equiv S$ and true otherwise. For $i \in S$, we let $p_i = 0$ almost surely, and for the null hypotheses we have$(p_j)_{j \notin S} \sim_{i.i.d.} \mathsf{Unif}([0,1])$.

Since $p_j = 0$ we also have $R_j = 1$ for all $j \in S$, and hence $W(j) \geq b_0$ for all $j \in S$. Consider a modified rule in which, every time a discovery is made, the potential is reset to b_0. Denote by $\widetilde{W}(j)$and $\widetilde{V}(j)$ the corresponding potential and number of false discoveries, respectively. Since the rule is monotone, we have $W(j) \geq \widetilde{W}(j)$ and hence $V(j) \geq \widetilde{V}(j)$, for all j. Further, for all $n \geq m_0$ we have $R(n) \geq 1$ and therefore

$$\mathrm{FDR}(n) = \mathbb{E}\Big\{\frac{V(n)}{R(n)}\Big\} = \mathbb{E}\Big\{\frac{V(n)}{\lfloor n/m_0 \rfloor + V(n)}\Big\} \geq \mathbb{E}\Big\{\frac{\widetilde{V}(n)}{\lfloor n/m_0 \rfloor + \widetilde{V}(n)}\Big\}, \tag{80}$$

where the last inequality follows since $x \mapsto x/(x+a)$ is monotone increasing for $x, a \geq 0$. Let $X_\ell(m_0)$, $\ell \geq 1$ denote the number of false discoveries (in the modified rule) between $H_{\ell m_0+1}$ and $H_{(\ell+1)m_0-1}$. Note that the $(X_\ell(m_0))_{\ell \geq 1}$ are mutually independent, bounded random variables and $\widetilde{V}(n) \geq \sum_{\ell=1}^{\lfloor n/m_0 \rfloor - 1} X_\ell(m_0)$. Hence, denoting by $X(m_0)$ an independent copy of the $X_\ell(m_0)$, we get

$$\liminf_{n\to\infty} \mathrm{FDR}(n) \geq \liminf_{n\to\infty} \mathbb{E}\Big\{\frac{\sum_{\ell=1}^{\lfloor n/m_0 \rfloor - 1} X_\ell(m_0)}{\lfloor n/m_0 \rfloor + \sum_{\ell=1}^{\lfloor n/m_0 \rfloor - 1} X_\ell(m_0)}\Big\} \tag{81}$$

$$= \frac{\mathbb{E}X(m_0)}{1 + \mathbb{E}X(m_0)}, \tag{82}$$

where the last equality follows from the strong law of large numbers and dominated convergence.

We can define $X(m_0)$ as the number of false discoveries under the modified rule between hypotheses H_1 and H_{m_0-1} when all nulls are true, i.e. $(p_j)_{j \geq 1} \sim_{i.i.d.} \mathsf{Unif}([0,1])$, and we initialize by $\widetilde{W}(0) = b_0$. By this construction, the sequence of random variables $(X(m_0))_{m_0 \geq 2}$ is monotone increasing with $\lim_{m_0\to\infty} X(m_0) = X(\infty)$, whence $\lim_{m_0\to\infty} \mathbb{E}X(m_0) = \mathbb{E}X(\infty)$ by monotone convergence. We next compute $\mathbb{E}X(\infty)$. Let T_1 be the time at which the first discovery is made (in particular, $\mathbb{P}(T_1 = \ell) = b_0\gamma_\ell \prod_{i=1}^{\ell-1}(1 - b_0\gamma_i)$). Denoting by $X(\ell, \infty) = \sum_{i=\ell+1}^{\infty} R_i$ the number of discoveries after time ℓ, we have

$$\mathbb{E}\{X(\infty)\} = \sum_{\ell=1}^{\infty} \mathbb{E}\{X(\infty)|T_1 = \ell\}\, \mathbb{P}(T_1 = \ell) \tag{83}$$

$$= \sum_{\ell=1}^{\infty} \mathbb{E}\{X(\infty)|T_1 = \ell\}\, \mathbb{P}(T_1 = \ell) \tag{84}$$

$$= \sum_{\ell=1}^{\infty} \mathbb{E}\{1 + X(\ell,\infty)|T_1 = \ell\}\, \mathbb{P}(T_1 = \ell) \tag{85}$$

$$= \sum_{\ell=1}^{\infty} \big\{1 + \mathbb{E}\{X(\infty)\}\big\}\, \mathbb{P}(T_1 = \ell) \tag{86}$$

$$= \big\{1 + \mathbb{E}\{X(\infty)\}\big\}\, \mathbb{P}(T_1 < \infty). \tag{87}$$

Note that by Eq. (82) we can assume $\mathbb{E}\{X(\infty)\} < \infty$. Since $\mathbb{P}(T_1 < \infty) = b_0$, the above implies $\mathbb{E}\{X(\infty)\} = b_0/(1 - b_0)$.

Substituting in Eq. (82), we deduce that, for any $\varepsilon > 0$, there exists $m_{0,*}(\varepsilon)$ such that, for the p-values constructed above with $m_0 \ge m_{0,*}(\varepsilon)$,

$$\liminf_{n\to\infty} \mathrm{FDR}(n) \ge (1-\varepsilon) b_0 \,. \tag{88}$$

Finally, we can take $\varepsilon \to 0$ if we modify the above construction by taking the set of non-null hypotheses to have, instead of equispaced elements, increasing gaps that diverge to infinity. For instance we can take the set of non-null to be $(H_{2^\ell})_{\ell\ge 2}$ and repeat the above analysis.

Equation (79) follows by the same argument. □

D FDR for dependent p-values: Proof of Theorem 3.6

Let $\mathcal{E}_{v,u}$ be the event that the generalized alpha-investing rule rejects exactly v true null and u false null hypotheses in $\mathcal{H}(n) = (H_1, \dots, H_n)$. We further denote by n_0 and $n_1 = n - n_0$ the number of true null and false null hypotheses in $\mathcal{H}(n)$. The false discovery rate for a fixed choice of the parameters $\boldsymbol{\theta}$ is

$$\begin{aligned} \mathrm{FDR}^\theta(n) &\equiv \mathbb{E}(\mathrm{FDP}^\theta(n)) & (89)\\ &= \sum_{v=0}^{n_0}\sum_{u=0}^{n_1} \frac{v}{(v+u)\vee 1}\,\mathbb{P}(\mathcal{E}_{v,u})\,. & (90)\end{aligned}$$

We next use a lemma from [BY01]. We present its proof here for the reader's convenience.

Lemma D.1 ([BY01])**.** *Let $\Omega_0 \subseteq [n]$ be the subset of true nulls. The following holds true:*

$$\mathbb{P}(\mathcal{E}_{v,u}) = \frac{1}{v}\sum_{i\in\Omega_0} \mathbb{P}((p_i \le \alpha_i) \cap \mathcal{E}_{v,u})\,. \tag{91}$$

Proof. Fix $\boldsymbol{\theta}$ and u, v. In particular $|\Omega_0| = n_0$. For a subset $\Omega \subseteq \Omega_0$ with $|\Omega| = v$, denote by $\mathcal{E}^{\Omega}_{v,u} \subseteq \mathcal{E}_{v,u}$ the event that the v true null hypotheses in Ω are the ones rejected, and additional u false null hypotheses are rejected.

Note that, for $i \in \Omega_0$, we have

$$\mathbb{P}\big((p_i \le \alpha_i) \cap \mathcal{E}^{\Omega}_{v,u}\big) = \begin{cases} \mathbb{P}(\mathcal{E}^{\Omega}_{v,u}) & \text{if } i \in \Omega\,,\\ 0 & \text{otherwise}\,. \end{cases} \tag{92}$$

Therefore,

$$\begin{aligned} \sum_{i\in\Omega_0} \mathbb{P}\big((p_i\le\alpha_i)\cap\mathcal{E}_{v,u}\big) &= \sum_{i\in\Omega_0}\sum_{\Omega\subseteq\Omega_0} \mathbb{P}\big((p_i\le\alpha_i)\cap\mathcal{E}^{\Omega}_{v,u}\big) & (93)\\ &= \sum_{\Omega\subseteq\Omega_0}\sum_{i\in\Omega_0} \mathbb{P}\big((p_i\le\alpha_i)\cap\mathcal{E}^{\Omega}_{v,u}\big) & (94)\\ &= \sum_{\Omega\subseteq\Omega_0}\sum_{i\in\Omega_0} \mathbb{I}(i\in\Omega)\,\mathbb{P}(\mathcal{E}^{\Omega}_{v,u}) = \sum_{\Omega\subseteq\Omega_0} v\,\mathbb{P}(\mathcal{E}^{\Omega}_{v,u}) = v\,\mathbb{P}(\mathcal{E}_{v,u})\,, & (95)\end{aligned}$$

which completes the proof. □

Applying Lemma D.1 in Eq. (90), we obtain

$$\mathrm{FDR}^{\theta}(n) = \sum_{v=0}^{n_0}\sum_{u=0}^{n_1} \frac{1}{(v+u)\vee 1} \sum_{i\in\Omega_0} \mathbb{P}((p_i \le \alpha_i) \cap \mathcal{E}_{v,u})\,. \tag{96}$$

Define the measure $\nu_{i,u,v}$ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ by letting, for any Borel set $A \in \mathcal{B}_{\mathbb{R}}$

$$\nu_{i,u,v}(A) \equiv \mathbb{P}\Big((p_i \le \alpha_i) \cap \mathcal{E}_{v,u} \cap \{\mathcal{I}_{i-1} \in A\}\Big) \tag{97}$$

Notice that, by definition, $\nu_{i,v,u}$ is supported on $[\mathcal{I}_{\min}(i-1), \mathcal{I}_{\max}(i-1)]$. Also $\nu_{i,v,u}$ is a finite measure, but not a probability measure (it does not integrate to one). Indeed $\int_{\mathcal{I}_{\min}(i-1)}^{\mathcal{I}_{\max}(i-1)} \mathrm{d}\nu_{i,v,u}(s) = \mathbb{P}((p_i \le \alpha_i) \cap \mathcal{E}_{v,u}\big)$. Then Eq. (96) yields

$$\mathrm{FDR}^{\theta}(n) = \sum_{i\in\Omega_0} \int_{\mathcal{I}_{\min}(i-1)}^{\mathcal{I}_{\max}(i-1)} \sum_{v=0}^{n_0}\sum_{u=0}^{n_1} \frac{1}{(v+u)\vee 1}\, \mathrm{d}\nu_{i,v,u}(s)\,. \tag{98}$$

Define $\nu_{i,k} = \sum_{v,u:v+u=k} \nu_{i,v,u}$. Note that, by definition of $R_i^{\mathrm{L}}(s)$, we have $\nu_{i,k}(\{s : k \le R_i^{\mathrm{L}}(s)\}) = 0$, whence $\nu_{i,k} = \mathbb{I}(k > R_i^{\mathrm{L}}(s))\nu_{i,k}$. Therefore:

$$\begin{aligned}
\mathrm{FDR}^{\theta}(n) &= \sum_{i\in\Omega_0} \int_{\mathcal{I}_{\min}(i-1)}^{\mathcal{I}_{\max}(i-1)} \sum_{k=R_i^{\mathrm{L}}(s)+1}^{n} \frac{1}{k}\, \mathrm{d}\nu_{i,k}(s) && (99)\\
&\le \sum_{i\in\Omega_0} \int_{\mathcal{I}_{\min}(i-1)}^{\mathcal{I}_{\max}(i-1)} \frac{1}{R_i^{\mathrm{L}}(s)+1} \sum_{k=R_i^{\mathrm{L}}(s)+1}^{n} \mathrm{d}\nu_{i,k}(s)\,. && (100)
\end{aligned}$$

Letting $\nu_i = \sum_{k=1}^{n} \nu_{i,k}$, we have, for any Borel set $A \in \mathcal{B}_{\mathbb{R}}$,

$$\begin{aligned}
\nu_i(A) &= \mathbb{P}\big(\{p_i \le \alpha_i\} \cap \{\mathcal{I}_{i-1} \in A\}\big) && (101)\\
&= \mathbb{P}\big(\{p_i \le g_i(\mathcal{I}_{i-1})\} \cap \{\mathcal{I}_{i-1} \in A\}\big) && (102)\\
&= \int_{\{\tau \le g_i(s)\}\cap\{s\in A\}} \mathrm{d}\widehat{\nu}_i(\tau, s)\,. && (103)
\end{aligned}$$

where $\widehat{\nu}_i$ is the joint probability measure of p_i and $\mathcal{I}_{i-1}$. Since g_i is non-decreasing and continuous, we will define its inverse by $g_i^{-1}(\tau) = \inf\{s : \, g_i(s) \geq \tau\}$. Using this in Eq. (100), we get the bound

$$
\begin{aligned}
&\mathrm{FDR}^{\theta}(n) \qquad\qquad (104)\\
&\leq \sum_{i \in \Omega_0} \int \frac{1}{R_i^{\mathrm{L}}(s)+1}\, \mathbb{I}\big(s \in [\mathcal{I}_{\min}(i-1), \mathcal{I}_{\max}(i-1)]\big)\, \mathbb{I}\big(\tau \in [0, g_i(s)]\big)\; \mathrm{d}\widehat{\nu}_i(\tau, s)\\
&\overset{(a)}{\leq} \sum_{i \in \Omega_0} \int \frac{1}{R_i^{\mathrm{L}}(s)+1}\, \mathbb{I}\big(s \in [\mathcal{I}_{\min}(i-1) \vee g_i^{-1}(\tau), \mathcal{I}_{\max}(i-1)]\big)\, \mathbb{I}\big(\tau \in [0, g_i(\mathcal{I}_{\max}(i-1))]\big)\;\; \mathrm{d}\widehat{\nu}_i(\tau, s)\\
&\overset{(b)}{\leq} \sum_{i \in \Omega_0} \int \Big\{ \frac{1}{R_i^{\mathrm{L}}(\mathcal{I}_{\min}(i-1))+1}\, \mathbb{I}\big(\tau \in [0, g_i(\mathcal{I}_{\min}(i-1))]\big)\\
&\qquad\qquad + \frac{1}{R_i^{\mathrm{L}}(g_i^{-1}(\tau))+1}\, \mathbb{I}\big(\tau \in [g_i(\mathcal{I}_{\min}(i-1)), g_i(\mathcal{I}_{\max}(i-1))]\big) \Big\}\; \mathrm{d}\widehat{\nu}_i(\tau, s)\\
&\overset{(c)}{\leq} \sum_{i \in \Omega_0} \Big\{ \frac{g_i(\mathcal{I}_{\min}(i-1))}{R_i^{\mathrm{L}}(\mathcal{I}_{\min}(i-1))+1} + \int \frac{1}{R_i^{\mathrm{L}}(g_i^{-1}(\tau))+1}\, \mathbb{I}\big(\tau \in [g_i(\mathcal{I}_{\min}(i-1)), g_i(\mathcal{I}_{\max}(i-1))]\big)\; \mathrm{d}\tau \Big\}
\end{aligned}
\tag{105}
$$

where (a) follows from monotonicity of $s \mapsto g_i(s)$, (b) by the monotonicity of $s \mapsto R_i^{\mathrm{L}}(s)$, and (c) follows by integrating over s and noting that $\mathrm{d}\widehat{\nu}_i(\tau)$ is the uniform (Lebesgue) measure on the interval $[0,1]$ since p_i is a p-value for a true null hypothesis. Therefore by the change of variables $\tau = g_i(s)$, we obtain

$$
\mathrm{FDR}^{\theta}(n) \leq \sum_{i \in \Omega_0} \Big\{ \frac{g_i(\mathcal{I}_{\min}(i-1))}{R_i^{\mathrm{L}}(\mathcal{I}_{\min}(i-1))+1} + \int_{\mathcal{I}_{\min}(i-1)}^{\mathcal{I}_{\max}(i-1)} \frac{\dot{g}_i(s)}{R_i^{\mathrm{L}}(s)+1} \mathrm{d}s \Big\}\,. \tag{106}
$$

Finally, let $\mathbf{0}_1^i$ be the zero sequence of length i. By definition $\mathcal{I}_i(\mathbf{0}_1^i) \geq \mathcal{I}_{\min}(i)$ and therefore, by definition of $R_i^{\mathrm{L}}(s)$ (26) we have $R_i^{\mathrm{L}}(\mathcal{I}_{\min}(i-1)) = 0$. The claim follows from equation (106).

E FDX control: Proof of Theorem 6.1

Define $u = (\gamma - b_0 - w_0)/(1-\alpha)$. We will denote by N the first time such that either $W(n) = 0$ or the condition in assumption G4 is violated, i.e.

$$
N \equiv \min\big\{\, n \geq 1 \;\text{ s.t. }\; W(n) = 0 \text{ or } M(n) + \xi_{n+1} > u \,\big\}\,. \tag{107}
$$

Note that this is a stopping time with respect to the filtration $\{\mathcal{F}_n\}$. Further, by assumption G4, there is no discovery after time N. Namely $R_j = 0$ for all $j > N$.

Define the process

$$
B(j) \equiv \begin{cases} b_0 R(j) - W(j) - V(j) + \gamma - b_0 & \text{if } j \leq N,\\ b_0 R(N) - W(N) - V(N) + \gamma - b_0 & \text{if } j > N. \end{cases} \tag{108}
$$

Note that $B(j)$ is measurable on $\mathcal{F}_j$. A key step will be to prove the following.

Lemma E.1. *The process $\{B(j)\}_{j\geq 0}$ is a submartingale with respect to the filtration $\{\mathcal{F}_j\}$.*

Proof. As already pointed out, $B(j)$ is measurable on $\mathcal{F}_j$. Further, since $\mathcal{F}_j$ is generated by j binary variables, $B(j)$ takes at most 2^j finite values, and is therefore integrable. Let $B'(j) \equiv b_0R(j) - W(j) - V(j) + \gamma - b_0$. Since B is a stopped version of B', it is sufficient to check that B' is a submartingale. Let $B'_j = B'(j) - B'(j-1)$, $W_j = W(j) - W(j-1)$, $V_j = V(j) - V(j-1)$. By definition, have $B'_j = b_0R_j - W_j - V_j = (b_0 - \psi_j)R_j - V_j + \varphi_j$.

We first assume that the null hypothesis H_j holds, i.e. $\theta_j = 0$. Hence, $R_j = V_j$ and $B'_j = (b_0 - \psi_j - 1)R_j + \varphi_j$. Taking conditional expectation, we get

$$\mathbb{E}(B'_j|\mathcal{F}_{j-1}) \overset{(a)}{=} (b_0 - \psi_j - 1)\mathbb{E}(R_j|\mathcal{F}_{j-1}) + \varphi_j \tag{109}$$

$$\overset{(b)}{\geq} (b_0 - \psi_j - 1)\alpha_j + \varphi_j \tag{110}$$

$$\overset{(c)}{\geq} (b_0 - \varphi_j/\alpha_j - b_0 + 1 - 1)\alpha_j + \varphi_j = 0\,, \tag{111}$$

where (a) follows because ψ_j and φ_j are measurable on $\mathcal{F}_{j-1}$; (b) by assumption (52), since $\psi_j \geq 0$ and $b_0 \leq 1$, whence $(b_0 - \psi_j - 1) \leq 0$, and (c) from assumption $\mathsf{G1}$, cf. Eq. (10).

Next, we assume a false null hypothesis, i.e. $\theta_j \neq 0$, and thus $V_j = 0$ and $B'_j = (b_0 - \psi_j)R_j + \varphi_j$. Taking again conditional expectation, we obtain

$$\mathbb{E}(B'_j|\mathcal{F}_{j-1}) = (b_0 - \psi_j)\mathbb{E}(R_j|\mathcal{F}_{j-1}) + \varphi_j \tag{112}$$

$$\geq (b_0 - \varphi_j - b_0)\mathbb{E}(R_j|\mathcal{F}_{j-1}) + \varphi_j \geq 0\,, \tag{113}$$

from assumption $\mathsf{G1}$, cf. Eq. (9). □

Applying Doob decomposition theorem, process $B(n)$ has a (unique) decomposition into a martingale $\widetilde{M}(n)$ and a nonnegative predictable process $A(n)$ that is almost surely increasing. Specifically,

$$\widetilde{M}(n) = B(0) + \sum_{j=1}^{n} \Big(B_j - \mathbb{E}(B_j|\mathcal{F}_{j-1})\Big)\,, \tag{114}$$

$$A(n) = \sum_{j=1}^{n} \mathbb{E}(B_j|\mathcal{F}_{j-1})\,. \tag{115}$$

We next define the process $Q(n)$ as follows.

$$Q(n) = \begin{cases} \widetilde{M}(n) & \text{if } \min_{0\leq i\leq n} \widetilde{M}(i) > 0\,, \\ 0 & \text{otherwise }. \end{cases} \tag{116}$$

Equivalently, define the stopping time

$$N_* \equiv \min\big\{\, n \geq 1 \;:\; \widetilde{M}(n) \leq 0 \,\big\}\,. \tag{117}$$

(Note that $\widetilde{M}(0) = B(0) = \gamma - b_0 - w_0 > 0$ by assumption $\mathsf{G3}$.) Then $Q(n)$ is the positive part of $\widetilde{M}(n)$ stopped at N_*:

$$Q(n) \equiv \max\big(\,0, \widetilde{M}(n \wedge N_*)\,\big)\,. \tag{118}$$

Since it is obtained by applying a convex function to a stopped submartingale, $Q(n)$ is also a submartingale.

Further,

$$0 \le Q(n) \le \max(\widetilde{M}(n), 0) \le \max_{n \le N} \widetilde{M}(n)\,, \tag{119}$$

where we used the fact that $\widetilde{M}(n)$ rains unchanged after N and $\widetilde{M}(0) > 0$. Observe that, defining $\widetilde{M}_j = \widetilde{M}(j) - \widetilde{M}(j-1)$, for $j \le N$,

$$\widetilde{M}_j = \begin{cases} (b_0 - \psi_j)(R_j - \mathbb{E}(R_j|\mathcal{F}_{j-1})) & \theta_j \neq 0\,, \\ (b_0 - \psi_j - 1)(R_j - \alpha_j) & \theta_j = 0\,. \end{cases} \tag{120}$$

Recalling definition of M_j, cf. Eq. (50), we have $\widetilde{M}_j \le M_j$ and $\widetilde{M}(0) = M(0) = \gamma - b_0 - w_0$. Hence $\widetilde{M}(n) \le M(n)$ for all $n \in \mathbb{N}$. Furthermore,

$$\max_{n \le N} M(n) = \max_{n \le N} \big(M(n-1) + M_n\big) \le \max_{n \le N} \big(u - \xi_n + M_n\big) \le u, \tag{121}$$

where we used the fact that $M(n-1) + \xi_n \le u$ by definition of stopping time N, cf. Eq. (107). Using (119), we obtain

$$0 \le Q(n) \le \max_{n \le N} M(n) \le u\,. \tag{122}$$

We next upper bound $\mathbb{P}(N_* < \infty)$ which directly yields an upper bound on FDX_γ. Define the event $\mathcal{E}_n \equiv \{Q(n) = 0\}$ and set $q_n \equiv \mathbb{P}(\mathcal{E}_n)$ for $n \in \mathbb{N}$. Using the sub-martingale property of $Q(n)$ and equation (122), we obtain

$$0 < \gamma - b_0 - w_0 = \mathbb{E}(Q(0)) \le \mathbb{E}(Q(n)) \le (1 - q_n)u\,, \tag{123}$$

whence we obtain $q_n \le \alpha$, by plugging in for u. Note that $\mathcal{E}_n \subseteq \mathcal{E}_{n+1}$ for all $n \in \mathbb{N}$. Clearly $\{N_* < \infty\} = \cup_{n=0}^{\infty} \mathcal{E}_n$ and by monotone convergence properties of probability measures

$$\mathbb{P}(N_* < \infty) = \lim_{n \to \infty} \mathbb{P}(\mathcal{E}_n) = \lim_{n \to \infty} q_n \le \alpha\,. \tag{124}$$

We lastly write FDX_γ in terms of event $\{N_* < \infty\}$ as follows.

$$\begin{aligned} \Big\{\sup_{n \ge 1} \mathrm{FDP}^\theta(n) \ge \gamma\Big\} &\overset{(a)}{\equiv} \Big\{\exists\, 1 \le n \le N : \, V^\theta(n) \ge \gamma(R(n) \vee 1)\Big\} \\ &\subseteq \Big\{\exists\, 1 \le j \le N : \, b_0 R(n) - V^\theta(n) + \gamma - b_0 \le 0\Big\} \\ &= \Big\{\exists\, 1 \le n \le N : \, B(n) \le -W(n)\Big\} \\ &\overset{(b)}{\subseteq} \Big\{\exists\, 1 \le n \le N : \, B(n) \le 0\Big\} \\ &\overset{(c)}{\subseteq} \Big\{\exists\, 1 \le n \le N : \, \widetilde{M}(n) \le 0\Big\} \\ &\subseteq \big\{N_* < \infty\big\}\,. \end{aligned} \tag{125}$$

Here (a) holds because there is no discovery after time N and $\mathrm{FDP}^\theta(n)$ remains unaltered; (b) holds since $W(n) \ge 0$ and (c) follows from the decomposition $B(n) = \widetilde{M}(n) + A(n)$ and $A(n) \ge 0$. Therefore,

$$\mathrm{FDX}_\gamma \le \mathbb{P}(N_* < \infty) \le \alpha\,. \tag{126}$$

F Proof of Theorem 4.1

We consider the following rule obtained by a modification of LORD :

$$\begin{aligned} W(0) &= b_0\,, \\ \varphi_i &= \alpha_i = b_0\gamma_{i-\tau_i}\,, \\ \psi_i &= b_0\,. \end{aligned} \tag{127}$$

In words, we replace $W(\tau_i)$ in φ_i with b_0. Given that in LORD , $W(\tau_i) \ge b_0$ at each step, the test level for rule (127) is smaller than or equal to the test level of LORD . Therefore, discoveries made by (127) are a subset of discoveries made by LORD and the statistical power of (127) lower bounds the power of LORD .

For testing rule (127), it is clear that the times between successive discoveries are i.i.d. under the mixture model. Therefore, the process $R(n) = \sum_{\ell=1}^{n} R_\ell$ is a renewal process. We let $\Delta_i = \tau_i - \tau_{i-1}$ be the i^{th} interval between discoveries and let $r_i \equiv \mathbb{I}(\tau_i \in \Omega_0^c)$ be the reward associated with inter-discovery Δ_i. In other words, at each discovery we get reward one only if that discovery corresponds to a non-null hypothesis. Recall that under the mixture model, each hypothesis is truly null/non-null independently of others. Therefore, (r_i, Δ_i) are i.i.d across index i and form a renewal-reward process. Clearly $\mathbb{E}(r_i) = \pi_1$ and we can compute $\mathbb{E}(\Delta_i)$ as follows:

$$\mathbb{P}(\Delta_i \ge m) = \mathbb{P}\Big(\cap_{\ell=1}^{m} \{p_\ell > \alpha_\ell\}\Big) = \prod_{\ell=1}^{m} \Big(1 - G(\alpha_\ell)\Big)\,. \tag{128}$$

Substituting for $\alpha_\ell = b_0\gamma_\ell$,

$$\mathbb{E}(\Delta_i) = \sum_{m=1}^{\infty} \mathbb{P}(\Delta_i \ge m) \le \sum_{m=1}^{\infty} \prod_{\ell=1}^{m} \Big(1 - G(b_0\gamma_\ell)\Big)\,. \tag{129}$$

Without loss of generality we can assume $\mathbb{E}(\Delta_i) < \infty$; otherwise bound (29) becomes trivial. Applying the strong law of large numbers for renewal-reward processes, the following holds true almost surely

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{R(n)} r_i = \frac{\mathbb{E}(r_i)}{\mathbb{E}(\Delta_1)} = \pi_1 \Big(\sum_{m=1}^{\infty} \prod_{\ell=1}^{m} \big(1 - G(b_0\gamma_\ell)\big)\Big)^{-1}\,. \tag{130}$$

Further, $\lim_{n\to\infty} |\Omega_0^c(n)|/n = \pi_1$, almost surely. Therefore, almost surely

$$\lim_{n\to\infty} \frac{1}{|\Omega_0^c(n)|} \sum_{i\in\Omega_0^c(n)} R_i \ge \lim_{n\to\infty} \frac{1}{|\Omega_0^c(n)|} \sum_{i=1}^{R(n)} r_i = \Big(\sum_{m=1}^{\infty} \prod_{\ell=1}^{m} \big(1 - G(b_0\gamma_\ell)\big)\Big)^{-1}\,, \tag{131}$$

where the first inequality follows from the fact that $\sum_{i=1}^{R(n)} r_i/|\Omega_0^c(n)|$ is the average power of rule (127), which as discussed, serves as a lower bound for the average power of LORD .

G Proof of Proposition 4.2

We set $\beta_m = b_0\gamma_m$ for $m \ge 1$. The Lagrangian for the optimization problem reads as

$$\mathcal{L} = \sum_{m=1}^{\infty} e^{-mG(\beta_m)} + \eta(\sum_{m=1}^{\infty} \beta_m - b_0)\,, \tag{132}$$

where η is a Lagrange multiplier. Setting the derivative with respect to β_m to zero, we get

$$\eta = mG'(\beta_m^{\rm opt})e^{-mG(\beta_m^{\rm opt})}\,. \tag{133}$$

Note that $G'(x) = \pi_1 F'(x) + (1-\pi_1) \ge 1-\pi_1$, since $F(x)$ is nondecreasing. Hence

$$m(1-\pi_1)e^{-mG(\beta_m^{\rm opt})} \le \eta\,,$$

whereby using the non-decreasing monotone property of G, we obtain

$$\beta_m^{\rm opt} \ge G^{-1}\Big(\frac{1}{m}\log\Big(\frac{m(1-\pi_1)}{\eta}\Big)\Big)\,. \tag{134}$$

To obtain the upper bound, note that be concavity of $F(x)$ on $(0,x_0)$, we have $G'(x) \le G(x)/x$ for $x \in (0,x_0)$. Since $\beta_m \to 0$ as $m \to \infty$, for large enough m, we have

$$mG(\beta_m^{\rm opt})e^{-mG(\beta^{\rm opt})} \ge \eta\beta_m^{\rm opt}\,. \tag{135}$$

Further, by equation (134) we have $mG(\beta_m^{\rm opt}) \ge 1$. Let $\xi_0 \ge 1$ be the solution of $\xi e^{-\xi} = \eta\beta_m^{\rm opt}$. Simple algebraic manipulation shows that $\xi_0 \le -2\log(\eta\beta_m^{\rm opt})$. Therefore, by Eq. (135) we have

$$mG(\beta_m^{\rm opt}) \le \xi_0 \le 2\log\Big(\frac{1}{\eta\beta_m^{\rm opt}}\Big) \le 2\log\Big(\frac{1}{\eta G^{-1}(1/m)}\Big)\,, \tag{136}$$

where the last step follows from Eq. (134). Using the non-decreasing property of $G(x)$ we get

$$\beta_m^{\rm opt} \le G^{-1}\Big(\frac{2}{m}\log\Big(\frac{1}{\eta G^{-1}(1/m)}\Big)\Big)\,, \tag{137}$$

The Lagrange multiplier η is chosen such that $\sum_{m=1}^{\infty}\beta_m^{\rm opt} = b_0$, equivalently, $\sum_{m=1}^{\infty}\gamma_m^{\rm opt} = 1$.

References

[AR14] Ehud Aharoni and Saharon Rosset, *Generalized α-investing: definitions, optimality results and application to public databases*, Journal of the Royal Statistical Society. Series B (Methodological) **76** (2014), no. 4, 771–794.

[BBM15] Mohsen Bayati, Sonia Bhaskar, and Andrea Montanari, *A low-cost method for multiple disease prediction*, AMIA Annual Symposium Proceedings, vol. 2015, American Medical Informatics Association, 2015, p. 329.

[BC15] Rina Foygel Barber and Emmanuel J. Candès, *Controlling the false discovery rate via knockoffs*, The Annals of Statistics **43** (2015), no. 5, 2055–2085.

[BC16] Rina Foygel Barber and Emmanuel J Candes, *A knockoff filter for high-dimensional selective inference*, arXiv:1602.03574 (2016).

[BH95] Yoav Benjamini and Yosef Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological) (1995), 289–300.

[Bic04] David R. Bickel, *On "strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates": Does a large number of tests obviate confidence intervals of the fdr?*, arXiv:q-bio/0404032 (2004).

[BR09] Gilles Blanchard and Étienne Roquain, *Adaptive false discovery rate control under independence and dependence*, J. Mach. Learn. Res. **10** (2009), 2837–2871.

[BSOM$^+$14] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar, *Big data in health care: using analytics to identify and manage high-risk and high-cost patients*, Health Affairs **33** (2014), no. 7, 1123–1131.

[BvdBS$^+$15] Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès, *Slopeadaptive variable selection via convex optimization*, The Annals of Applied Statistics **9** (2015), no. 3, 1103.

[BY01] Yoav Benjamini and Daniel Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, Annals of statistics (2001), 1165–1188.

[DFH$^+$14] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth, *Preserving statistical validity in adaptive data analysis*, arXiv:1411.2664, 2014.

[DJ94] David L. Donoho and Iain M. Johnstone, *Minimax risk over l_p balls*, Prob. Th. and Rel. Fields **99** (1994), 277–303.

[FS07] Dean P. Foster and Robert A. Stine, *Alpha-investing: A procedure for sequential control of expected false discoveries*, http://gosset.wharton.upenn.edu/research/edc.pdf, 2007.

[FST14] William Fithian, Dennis Sun, and Jonathan Taylor, *Optimal inference after model selection*, arXiv:1410.2597 (2014).

[GLN02] Christopher R Genovese, Nicole A Lazar, and Thomas Nichols, *Thresholding of statistical maps in functional neuroimaging using the false discovery rate*, Neuroimage **15** (2002), no. 4, 870–878.

[GW06] Christopher R Genovese and Larry Wasserman, *Exceedance control of the false discovery proportion*, Journal of the American Statistical Association **101** (2006), no. 476, 1408–1417.

[GWCT15] Max Grazier G'Sell, Stefan Wager, Alexandra Chouldechova, and Robert Tibshirani, *Sequential selection procedures and false discovery rate control*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) (2015).

[Ioa05a] John PA Ioannidis, *Contradicted and initially stronger effects in highly cited clinical research*, Jornal of the American Medical Association **294** (2005), no. 2, 218–228.

[Ioa05b] ———, *Why most published research findings are false*, PLoS medicine **2** (2005), no. 8, e124.

[JC07] Jiashun Jin and T Tony Cai, *Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons*, Journal of the American Statistical Association **102** (2007), no. 478, 495–506.

[Jin08] Jiashun Jin, *Proportion of non-zero normal means: universal oracle equivalences and uniformly consistent estimators*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70** (2008), no. 3, 461–493.

[JM13] Adel Javanmard and Aandrea Montanari, *Nearly optimal sample size in hypothesis testing for high-dimensional regression*, 51st Annual Allerton Conference (Monticello, IL), June 2013, pp. 1427–1434.

[JM14a] Adel Javanmard and Andrea Montanari, *Confidence intervals and hypothesis testing for high-dimensional regression*, The Journal of Machine Learning Research **15** (2014), no. 1, 2869–2909.

[JM14b] ———, *Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory*, IEEE Trans. on Inform. Theory **60** (2014), no. 10, 6522–6554.

[JM15] ———, *On online control of false discovery rate*, arXiv:1502.06197 (2015).

[Joh94] I.M. Johnstone, *On minimax estimation of a sparse normal mean vector*, Annals of Statistics **22** (1994), 271289.

[LB16] Ang Li and Rina Foygel Barber, *Accumulation tests for fdr control in ordered hypothesis testing*, Journal of the American Statistical Association (2016), no. just-accepted, 1–38.

[LFU11] Dongyu Lin, Dean P Foster, and Lyle H Ungar, *Vif regression: A fast regression algorithm for large data*, Journal of the American Statistical Association **106** (2011), no. 493, 232–247.

[LR12] Erich Leo Lehmann and Joseph P Romano, *Generalizations of the familywise error rate*, Springer, 2012.

[LTTT14] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani, *A significance test for the lasso*, Annals of statistics **42** (2014), no. 2, 413.

[MR06] Nicolai Meinshausen and John Rice, *Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses*, The Annals of Statistics **34** (2006), no. 1, 373–393.

[Owe05] Art B Owen, *Variance of the number of false discoveries*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 3, 411–426.

[PSA11] Florian Prinz, Thomas Schlange, and Khusru Asadullah, *Believe it or not: how much can we rely on published data on potential drug targets?*, Nature reviews Drug discovery **10** (2011), no. 9, 712–712.

[PWJ15] Leo Pekelis, David Walsh, and Ramesh Johari, *The new stats engine*, `http://pages.optimizely.com/rs/optimizely/images/stats_engine_technical_paper.pdf`, 2015.

[RAN14] Saharon Rosset, Ehud Aharoni, and Hani Neuvirth, *Novel statistical tools for management of public databases facilitate community-wide replicability and control of false discovery*, Genetic epidemiology **38** (2014), no. 5, 477–481.

[RYB03] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini, *Identifying differentially expressed genes using false discovery rate controlling procedures*, Bioinformatics **19** (2003), no. 3, 368–375.

[ST03] John D Storey and Robert Tibshirani, *Sam thresholding and false discovery rates for detecting differential gene expression in dna microarrays*, The analysis of gene expression data, Springer, 2003, pp. 272–290.

[Sto02] John D Storey, *A direct approach to false discovery rates*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **64** (2002), no. 3, 479–498.

[TLTT14] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani, *Exact post-selection inference for forward stepwise and least angle regression*, arXiv:1401.3889 (2014).

[VdGBRD14] Sara Van de Geer, Peter Bühlmann, Yaacov Ritov, and Ruben Dezeure, *On asymptotically optimal confidence regions and tests for high-dimensional models*, The Annals of Statistics **42** (2014), no. 3, 1166–1202.

[vdLDP04] Mark J van der Laan, Sandrine Dudoit, and Katherine S Pollard, *Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives*, Statistical applications in genetics and molecular biology **3** (2004), no. 1.

[ZZ14] Cun-Hui Zhang and Stephanie S Zhang, *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **76** (2014), no. 1, 217–242.