Electronic Journal of Statistics
Vol. 11 (2017) 1983–2001
ISSN: 1935-7524
DOI: 10.1214/17-EJS1277

Distribution-free multiple testing

Ery Arias-Castro and Shiyun Chen

Department of Mathematics, University of California, San Diego, USA
e-mail: eariasca@ucsd.edu*;* shc176@ucsd.edu

Abstract: We study a stylized multiple testing problem where the test statistics are independent and assumed to have the same distribution under their respective null hypotheses. We first show that, in the normal means model where the test statistics are normal Z-scores, the well-known method of Benjamini and Hochberg [4] is optimal in some asymptotic sense. We then show that this is also the case of a recent distribution-free method proposed by Barber and Candès [14]. The method is distribution-free in the sense that it is agnostic to the null distribution — it only requires that the null distribution be symmetric. We extend these optimality results to other location models with a base distribution having fast-decaying tails.

MSC 2010 subject classifications: 62G10, 62G20.
Keywords and phrases: Multiple testing, distribution-free procedure, Benjamini-Hochberg procedure, asymptotic optimality, false discovery rate (FDR) control.

Received May 2016.

Contents

1. Introduction

Multiple testing arises in a wide array of applied settings, ranging from anomaly detection in sensor arrays to the selection of genes that are differentially expressed [8, 10]. This is particularly true in so-called discovery science, where the scientist proceeds by formulating hypotheses, testing each one of them on data, and following up on the most promising ones. Each step along the way is fraught with pitfalls, and even if the experiment was correctly designed and carried out, the scientist still needs to contend with the multitude of tests that were performed.

Multiple testing is now a well-established area in statistics. In a substantial proportion of the corresponding literature it is assumed that P-values are available. This, implicitly, assumes that the null distribution of each test statistic is known (perfectly). For example, the Benjamini-Hochberg (BH) procedure was proposed in this context [4]. See [27] for a fairly recent and comprehensive review of the literature, as it pertains to mathematical results in the area.

Our contribution is two-fold. First, we prove that the BH method is asymptotically optimal to first order in the normal (location) model, which corresponds to an idealized setting where the tests being performed are Z-tests and the effect, when present, affects the mean. In fact, we show that this is the case in the much wider context of asymptotically generalized Gaussian models — see Definition 1. Second, we propose to use the recent distribution-free method of Barber and Candès [14] that only relies on the assumption that the test statistics have a common null distribution that is symmetric about 0 and show that, in the same normal model, it achieves the same asymptotic performance to first order. This method, proposed in the context of post-model selection inference, is also intimately related to our own work [2] on distribution-free testing of the global null hypothesis.

1.1. The risk of a multiple testing procedure

Consider a setting where we want to test n null hypotheses, denoted $\mathbb{H}_1, \dots, \mathbb{H}_n$. The test that we use for $\mathbb{H}_i$ rejects for large positive values of a statistic X_i. Throughout, we assume that $X_1, \dots, X_n$ are independent. Denote the vector of test statistics by $\mathbf{X} = (X_1, \dots, X_n)$. Let Ψ_i denote the survival function[1] of X_i and $\mathbf{\Psi} = (\Psi_1, \dots, \Psi_n)$.

Remark 1. In a large portion of the literature, it is assumed that P-values can be computed (or at least approximated). The simplest such case is when $\mathbb{H}_i$ is a singleton, $\mathbb{H}_i = \{\Psi_i^{\text{null}}\}$, and the null distributions $\Psi_1^{\text{null}}, \dots, \Psi_n^{\text{null}}$ are known. In that case, the i-th P-value is defined as $P_i = \Psi_i^{\text{null}}(X_i)$, which is the probability

[1] In this paper, the survival function of a random variable Y is defined as $y \mapsto \mathbb{P}(Y \geq y)$.

of exceeding the observed value of the statistic under its null distribution. In this context, working with the statistics $X_1, \dots, X_n$ is equivalent to working with the P-values $P_1, \dots, P_n$.

Let $\mathcal{F} \subset [n] := \{1, \dots, n\}$ index the false null hypotheses, meaning

$$\mathcal{F} = \{i \in [n] : \Psi_i \notin \mathbb{H}_i\}. \tag{1}$$

A multiple testing procedure $\mathcal{R}$ takes the test statistics $\mathbf{X}$ and return a subset of $\mathcal{R}(\mathbf{X}) \subset [n]$ representing the null hypotheses that the procedure rejects. Given such a procedure $\mathcal{R}$, the false discovery rate is defined as the expected value of the false discovery proportion in [4]

$$\textsc{fdr}_{\mathbf{\Psi}}(\mathcal{R}) = \mathbb{E}_{\mathbf{\Psi}}(\textsc{fdp}(\mathcal{R}(\mathbf{X}))), \quad \textsc{fdp}(\mathcal{R}) := \frac{|\mathcal{R} \setminus \mathcal{F}|}{|\mathcal{R}|}, \tag{2}$$

where we denoted the cardinality of a set $\mathcal{A} \subset [n]$ by $|\mathcal{A}|$ and with the convention that $0/0 = 0$. While the FDR of a multiple testing procedure is analogous to the level or size of a test procedure, the false non-discovery rate (FNR) plays the role of power and is defined as the expected value of the false non-discovery proportion[2]

$$\textsc{fnr}_{\mathbf{\Psi}}(\mathcal{R}) = \mathbb{E}_{\mathbf{\Psi}}(\textsc{fnp}(\mathcal{R}(\mathbf{X}))), \quad \textsc{fnp}(\mathcal{R}) := \frac{|\mathcal{F} \setminus \mathcal{R}|}{|\mathcal{F}|}. \tag{3}$$

In analogy with the risk of a test — which is defined as the sum of the probabilities of type I and type II error — we define the risk of a multiple testing procedure $\mathcal{R}$ as the sum of the false discovery rate and the false non-discovery rate

$$\text{risk}_{\mathbf{\Psi}}(\mathcal{R}) = \textsc{fdr}_{\mathbf{\Psi}}(\mathcal{R}) + \textsc{fnr}_{\mathbf{\Psi}}(\mathcal{R}). \tag{4}$$

Remark 2. The procedure that never rejects and the one that always reject both achieve a risk of 1, so that any method that has a risk exceeding 1 is useless.

1.2. Threshold procedures

We say that a multiple testing procedure $\mathcal{R}$ is of threshold type if it is of the form

$$\mathcal{R}(X_1, \dots, X_n) = \{i : X_i \geq \tau(X_1, \dots, X_n)\}, \tag{5}$$

for some threshold function τ. For example, the BH method is a threshold procedure based on the P-values — see (11).

Because they are so natural in the present context, we will restrict the discussion to threshold procedures. In particular, the lower bound that we develop (Theorem 1) is only meant to apply to such procedures.

A sizable proportion of the papers in the literature do the same — see [27]. This is for example the case of [28].

[2] This definition is different from that of [16].

1.3. The normal model and the optimality of the BH method

This model corresponds to the setting above with $X_i \sim \Psi_i = \mathcal{N}(\mu_i, 1)$ and $\mathbb{H}_i : \mu_i = 0$, so that $\mathbb{H}_i$ is a singleton equal to $\Psi_i^{\rm null} = \mathcal{N}(0, 1)$. In this context it is compelling to ask how large the μ_i's need to be in order for the risk of the BH procedure to tend to zero. To the best of our knowledge, this question has not been directly answered in the literature.

Our inspiration for considering the normal (location) model comes from the seminal work of Ingster [18, 19] and [9] on testing the global null $\bigcap_i \mathbb{H}_i$. In [18] we find the following first-order asymptotic result. Assume a prior under which $m \le n$ randomly picked μ_i's are set to $\sqrt{2r \log n}$ and the others are set to 0. An interesting parameterization happens to be $m/n \sim n^{-\beta}$ with $\beta > 0$ fixed. Focusing on the so-called sparse regime, where $\beta > 1/2$, one finds that the detection boundary is at $r = \rho(\beta)$, where

$$\rho(\beta) = \begin{cases} \beta - 1/2, & 1/2 < \beta \le 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1. \end{cases} \tag{6}$$

This means that, taking r to be fixed, when $r < \rho(\beta)$ all tests have risk at least 1 in the large sample limit (which is as bad as random guessing), while when $r > \rho(\beta)$ the likelihood ratio test has risk 0 in the large sample limit. Donoho and Jin [9] propose an adaptive test procedure based on Tukey's higher criticism that achieves this optimal detection boundary. (The higher criticism also achieves the detection boundary over $\beta \le 1/2$ not displayed here.)

Returning to the question of identifying the false null hypotheses, which is our concern here, we know that $r > 1$ allows for the identification of the false nulls with a control of the family-wise error rate (FWER) at any fixed level. In fact, if we define the corresponding risk as the sum of FWER and the probability of at least one false non-discovery, then $r = 1$ is the precise boundary for this to be controlled, and the Bonferroni procedure achieves the boundary over $\beta \in (0, 1)$ — we leave this as an exercise to the reader. In this paper, we focus instead on controlling the risk (4) involving FDR. The following is a special case of a more general lower bound appearing later in the paper.

Corollary 1. *In the normal model, assume that $\beta \in (0, 1)$ and $r \ge 0$ are both fixed. If $r < \beta$, then the risk of any threshold procedure has limit inferior at least 1 as $n \to \infty$.*

In our context, we know that Corollary 1 is tight because the BH method (which is a threshold procedure) achieves the stated selection boundary with FDR control level set at some $q \to 0$ slowly. The following is also a special case of a more general result appearing later in the paper.

Corollary 2. *In the setting of Corollary 1, if instead $r > \beta$, then the risk of the BH procedure (properly calibrated) tends to 0 as $n \to \infty$.*

It is worth remembering that BH method is known to control the FDR at the prescribed level [4], so the result is really about its (asymptotic) control of the FNR.

Together, Corollary 1 and Corollary 2 establish the BH procedure as asymptotically optimal to first order in the normal model among threshold procedures. We will see that this remains true for a much wider class of models.

Remark 3. While the equation $r = \rho(\beta)$ defines the "detection boundary" in the (β, r) plane for testing the global null $\bigcap_i \mathbb{H}_i$ in the normal model, the equation $r = \beta$ is the "selection boundary" for the same multiple testing problem. Intuitively, detection is "easier" than selection, and this is confirmed in the fact that the detection boundary is entirely below the selection boundary — indeed, $\beta > \rho(\beta)$ for all $\beta \in (0, 1)$.

1.4. Multiple testing under symmetry

The P-values are based on the assumed knowledge of the null distribution of each test statistic. In many practical settings, this is not strictly the case, resulting in P-values that are only approximately uniformly distributed under their respective null hypothesis. This can jeopardize the control of the FDR. In the same way that it may be appealing in some situations to use a distribution-free test such as the signed-rank test instead of the t-test, it may also be desirable to use a distribution-free procedure for multiple testing.

Our working assumption is the following

- *$X_1, \dots, X_n$ are independent with common null distribution that is symmetric about 0.*

This assumption might be reasonable in some crossover trials. Although testing the global null is more typical in such a setting (and one might apply the signed-rank test), a proper multiple testing analysis may be carried out when it is desired to identify which subjects truly benefited from treatment.

The assumption of symmetry is at the very core of the literature on nonparametric tests [17]. And it is also quite natural in the context of multiple testing. For example, under these assumptions, [11] consider testing the global null and propose a test based on sign flips, while [2] propose a nonparametric analog to the higher criticism. Beyond testing the global null, [3] propose a resampling procedure also based on sign flips with the purpose of controlling the FWER in a setting that also allows for dependence, while [14] propose a nonparametric analog to the BH method.

We call the latter the Barber-Candès (BC) procedure — see Section 4 for a proper definition. We study this method and show that, under fairly general conditions, it achieves the selection boundary. In particular, it does as well as the BH procedure which requires the knowledge of the null distributions. The following is a special case of a more general results appearing later on.

Corollary 3. *The conclusions of Corollary 2 apply to the BC procedure.*

The BC method is shown in [14] to control the FDR at the desired level, so the result is really about its (asymptotic) control of the FNR.

1.5. More related work

Our contribution is thus twofold: we obtain an asymptotic oracle risk bound for multiple testing and then show that the BH achieves that bound; and we show that the distribution-free BC method also achieves that bound.

Various oracle bounds are available in the literature. In a context where the P-values are uniformly distributed under the null and have the same distribution under the alternative, [16] consider an oracle that knows the number of false null hypotheses $|\mathcal{F}|$ and the common alternative distribution. See also [5, 24, 28, 29]. [23] also discusses oracle bounds but in a different setting where FWER control is the goal.

The notion of risk considered here (4), although natural to us, seems new. More common is the risk corresponding to Hamming loss, very popular in the classification literature. In our notation, for a procedure $\mathcal{R}$, this risk is defined as follows

$$\text{risk}_{\boldsymbol{\Psi}}^{\text{Hamming}}(\mathcal{R}) = \mathbb{E}_{\boldsymbol{\Psi}}(|\mathcal{R} \triangle \mathcal{F}|). \tag{7}$$

For example, this risk is considered in [5, 6, 16, 20, 22, 24, 29]. All these papers provide some asymptotic analysis of the Hamming risk, whether from a minimax or oracle perspective. In this context, [5, 16, 24] compare the performance of BH method to that of an oracle, concluding that the BH method comes close to achieving the oracle bound under some conditions.

Other distribution-free procedures have been suggested in the literature. Most are based on resampling [3, 15, 26, 30, 31]. These methods are not applicable in the setting assumed here. They are typically applied to situations, as in microarray analysis, where each test statistic is based on comparing two (or more) samples. Another class of methods consist in estimating the null distribution — assumed common to all test statistics — and the alternative distribution — also assumed to be common to all test statistics — with the goal of imitating the oracle thresholding method based on that knowledge. This is advocated in [12, 25], for example. [28] and [29] discuss such procedures and derive performance bounds. Such methods rely on the ability to estimate the mixture consistently. There is work in that direction in [7, 21].

Although not as directly related, [1] consider the problem of estimating the mean vector $(\mu_1, \ldots, \mu_n)$ in the normal model, and show that hard thresholding with the BH threshold is asymptotically minimax in some settings.

1.6. Content

In Section 2 we derive an oracle bound on the boundary for multiple testing in a location model where the base distribution is asymptotically generalized Gaussian. This comprises the normal model. In Section 3 we analyze the performance of the BH procedure based on the full knowledge of the null distribution, while in Section 4 we analyze the performance of the BC procedure. We present the result of some numerical experiments in Section 5. The proofs are gathered in Section 6.

2. The AGG model

We start by defining an oracle threshold procedure, which will serve as benchmark on a family of location models where the base distribution is asymptotically polynomial in log-scale — which in particular encompasses the normal model. The result is an oracle risk bound.

2.1. The oracle procedure

We consider an oracle that provides $\mathcal{F}$ and use that information to optimize the threshold in terms of minimizing the risk at a particular realization, namely,

$$\tau_o(\mathbf{X}) \in \arg\min_{t\in\mathbb{R}} \ \textsc{fdp}(\mathcal{R}_t(\mathbf{X})) + \textsc{fnp}(\mathcal{R}_t(\mathbf{X})), \quad \mathcal{R}_t(\mathbf{X}) := \{i : X_i \geq t\}. \qquad (8)$$

In words, with full knowledge of the set of false null distributions $\mathcal{F}$, the procedure chooses a threshold that partitions the test statistics in a way that minimizes the sum of the false discovery and non-discovery proportions. The expected risk of this procedure is what we call below the oracle risk.

Remark 4. Of course, if one knew $\mathcal{F}$, one would simply reject $\mathbb{H}_i$ for all $i \in \mathcal{F}$ and, in the end, there would not any multiple testing problem to deal with! The oracle procedure is, however, constrained to be of threshold type, with the goal of serving as a benchmark for threshold-type procedures.

Our oracle is the strongest possible, in the sense that it provides $\mathcal{F}$, and we use the oracle information to optimize the threshold. Most other publications that discuss oracle bounds, such as [5, 16, 24, 28, 29], operate in a setting where the statistics have the same null distribution and the same alternative distribution, and consider an oracle that provides these two distributions together with the number of false null hypotheses $|\mathcal{F}|$; this oracle information is then used to optimize a *constant* threshold. [23] consider an oracle that provides $\mathcal{F}$, and well as the joint distribution of the P-values indexed by $\mathcal{F}^c$, and use that oracle information to obtain an optimized single-step procedure for FWER control.

2.2. Asymptotically generalized Gaussian model

In a location model, we assume that we know the null survival function Ψ, assumed to be continuous for simplicity, and consider $\Psi(\cdot - \mu)$ as a location family of distributions. We then assume that the test statistics are independent with respective distribution $X_i \sim \Psi_i = \Psi(\cdot - \mu_i)$, where $\mu_i = 0$ under the null $\mathbb{H}_i$ and $\mu_i > 0$ otherwise. Both minimax and Bayesian considerations lead to considering a prior on the μ_i's where $m \leq n$ randomly picked μ_i's are set equal to some $\mu > 0$ and the others are set to 0. The prior is therefore defined based on m and μ, which together control the signal strength.

Beyond the normal model, we consider other location models where the base distribution has a polynomial right tail in log scale.

Definition 1. A survival function Ψ is asymptotically generalized Gaussian (AGG) on the right with exponent $\gamma > 0$ if $\lim_{x\to\infty} x^{-\gamma} \log \Psi(x) = -1/\gamma$.

The AGG class of distributions is nonparametric and quite general. It includes the parametric class of generalized Gaussian (GG) distributions with densities $\{\psi_\gamma, \gamma > 0\}$ given by $\log \psi_\gamma(x) \propto -|x|^\gamma/\gamma$, which comprises the normal distribution ($\gamma = 2$) and the double exponential distribution ($\gamma = 1$). We assume that $\gamma \geq 1$ so that the null distribution has indeed a sub-exponential right tail.

Remark 5. We note that the scale (e.g., standard deviation) is fixed, but this is really without loss of generality as both the BH and BC methods are scale invariant. For the BH method, this is because the P-values are scale invariant. However, this is so because we provide the BH method with the null distribution, including the scale. The BC method, by contrast, can operate without knowledge of the scale.

Donoho and Jin [9] consider the problem of testing the global null in a GG location model and derived the detection boundary. We use the same prior, where m nulls chosen uniformly at random are designated to be false and all positive μ_i's are set equal to μ, with

$$m = \lfloor n^{1-\beta} \rfloor, \qquad \text{with } 0 < \beta < 1 \quad \text{(fixed)}, \tag{9}$$

and

$$\mu = \mu_\gamma(r) = (\gamma r \log n)^{1/\gamma}, \qquad \text{with } r > 0 \quad \text{(fixed)}. \tag{10}$$

Neuvial and Roquain [24] obtain general bounds on the excess (Hamming) risks of Bayesian FDR and the BH method relative to an oracle, which they specialize to the GG model, showing that under similar conditions the BH method achieves an oracle bound.

Theorem 1. *Consider a location model where the base distribution is AGG with exponent $\gamma \geq 1$, with prior described above, and with the parameterization (9)-(10). If $r < \beta$, then the oracle risk has limit inferior at least 1 as $n \to \infty$.*

3. The performance of the BH method

We order the X_i's in *decreasing* order, to obtain the following order statistics $X_{(1)} \geq \cdots \geq X_{(n)}$. Given a desired FDR control at q, the BH procedure of [4] is defined as the threshold procedure (5), with threshold

$$\tau_{\rm BH} = X_{(\iota_{\rm BH})}, \quad \iota_{\rm BH} := \max\left\{i : X_{(i)} \geq \Psi^{-1}(iq/n)\right\}. \tag{11}$$

This procedure is shown in [4] to control the FDR at q when the tests are independent — which we assume throughout.

Typically, q is set to a small number, like $q = 0.10$. In this paper we allow $q \to 0$ as $n \to \infty$, but slowly. Specifically, we always assume that

$$q = q(n) > 0 \text{ such that } n^a q(n) \to \infty \text{ for all fixed } a > 0. \tag{12}$$

The following result establishes the BH procedure as optimal in the AGG model, in the sense that it achieves the selection boundary ($r = \beta$) stated in Theorem 1.

Theorem 2. *In the setting of Theorem 1, if instead $r > \beta$, then the BH procedure with q satisfying (12) has* FNR *tending to 0 as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0 since the procedure has* FDR $\leq q$*.*

Remark 6. For any multiple testing procedure, FNR $\to 0$ if and only if FNP $\to 0$ in probability. Indeed, one direction is justified by Markov's inequality, and the other direction is justified by dominated convergence and the fact that FNP ≤ 1.

4. The performance of the BC method

Under the assumption of symmetry, given the desired FDR control level q, the Barber-Candès (BC) procedure defines the data-dependent threshold τ_{BC} as:

$$\tau_{\text{BC}} = \inf \left\{ t \in |\mathbf{X}| : \widehat{\text{FDP}}(t) \leq q \right\}, \tag{13}$$

where, as usual, the infimum is infinite if the set is empty, $|\mathbf{X}| := \{|X_i| : i = 1, \dots, n\}$ is the set of sample absolute values, and

$$\widehat{\text{FDP}}(t) := \frac{1 + \#\{i : X_i \leq -t\}}{1 \vee \#\{i : X_i \geq t\}}, \tag{14}$$

is a measure of how asymmetric the set of observations $\{X_i : |X_i| \geq t\}$ is.

The notation is borrowed from [14] and is justified by the fact that this quantity aims at estimating FDP$(\mathcal{R}_t)$, where $\mathcal{R}_t = \{i : X_i \geq t\}$ as in (8). The BC procedure is shown in [14] to control the FDR at level q.

The following result shows that, although agnostic to the null distribution, the BC procedure achieves the selection boundary in a AGG model as long as the underlying distribution is symmetric.

Theorem 3. *In the setting of Theorem 1, and assuming that the null distribution Ψ is symmetric about 0, if instead $r > \beta$, then the BC procedure with q satisfying (12) has* FNR *tending to 0 as $n \to \infty$. In particular, if $q \to 0$, then it has risk tending to 0 since the procedure has* FDR $\leq q$*.*

5. Numerical experiments

In this section, we perform simple simulations to compare the BH and BC procedures on finite data, with the goal of illustrating the theory we established. We consider the normal model and the double-exponential model. It is worth repeating that the BH procedure requires knowledge of null distribution as it is based on the P-values. In contrast, the BC method does not require knowledge of the null distribution.

5.1. Fixed sample size

In this first set of experiments, the sample size is chosen large at $n = 10^5$. The FDR control level is set at $q = 0.05$. We draw m observations from the alternative distribution $\Psi(\cdot - \mu)$, and the other $n - m$ from the null distribution Ψ. All the models are parameterized as described in Section 2.2, in particular, (9) and (10). We choose a few values for the parameter β so as to exhibit different sparsity levels, while the parameter r takes values in a grid of spanning $[0, 1]$. Each situation is repeated 500 times and we report the average FDP and FNP for each procedure.

5.1.1. Normal model

In this model Ψ is the standard normal distribution. The simulation results are reported in Figure 1 and Figure 2. In Figure 1 we report the FDP. Recall that the methods are set to control the FDR at the desired level ($q = 0.05$). We see that the BC method becomes more conservative than the BH method as β increases. In Figure 2 we report the FNP. We see that the BC method performs comparably to the BH method at $\beta = 0.3$ and $\beta = 0.5$, but is clearly less powerful in the sparsest regime $\beta = 0.7$. This is in line with the earlier observation that the BC method becomes more conservative with increasing values of β. It can also be explained by the fact, at $\beta = 0.7$, the number of false nulls ($m = 31$ out of $n = 10^5$) is too small to reveal the asymptotic power of the BC method. Finally, we remark that the transition from high FNP to low FNP happens in the vicinity of the theoretical threshold ($r = \beta$).

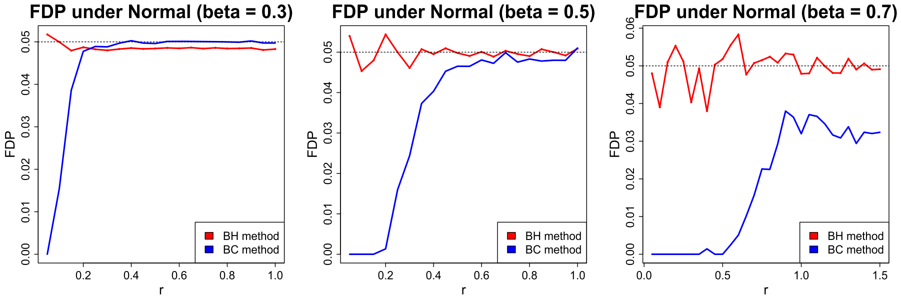

Fig 1. *Simulation results showing the FDP for the BH and BC methods under the normal model in three distinct sparsity regimes. The black horizontal line delineates the desired FDR control level ($q = 0.05$).*

5.1.2. Double-exponential model

In this model Ψ is double-exponential distribution with variance of 1. The simulation results are reported in Figure 3 (FDP) and Figure 4 (FNP). Here we observe that the BC method is rather conservative regardless of β. The two methods are again comparable in terms of FNP, in fact a bit more so than in

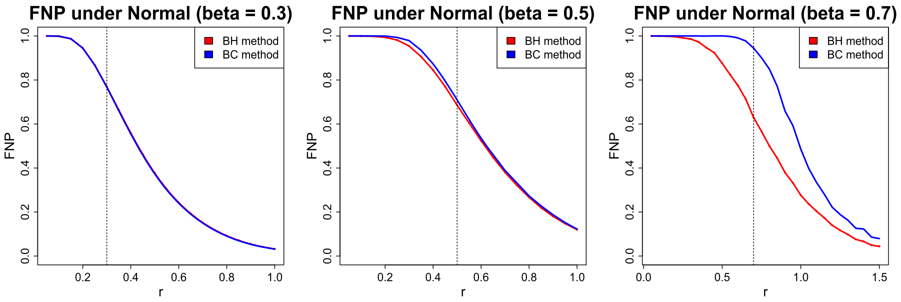

Fig 2. *Simulation results showing the FNP for the BH and BC methods under the normal model in three distinct sparsity regimes. The black vertical line delineates the theoretical threshold* ($r = \beta$)*.*

the normal setting. The transition from FNP near 1 to FNP near 0 happens, again, in the vicinity of the theoretical threshold, but is much sharper here.

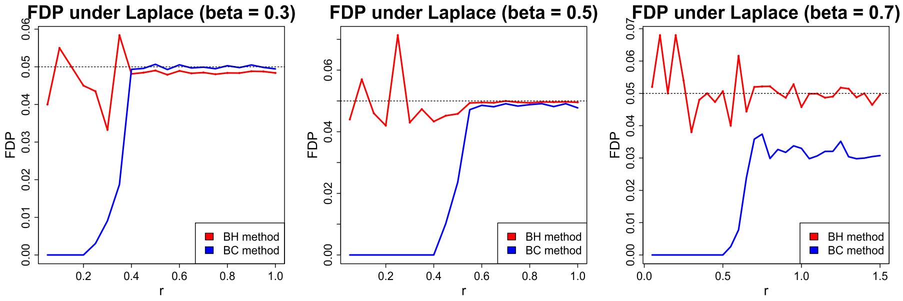

Fig 3. *Simulation results showing the FDP for the BH and BC methods under the double-exponential model in three distinct sparsity regimes. The black horizontal line delineates the desired FDR control level* ($q = 0.05$)*.*

5.2. Varying sample size

In this second set of experiments, we examine the effect of various sample sizes on the risk of BH and BC procedures under the standard normal model and the double-exponential model (with variance 1). We simultaneously explore the effect of letting the desired FDR control level q tend to 0, in accordance with (12). Specifically, we set it as $q = q_n = 1/\log n$. We choose n on a log scale, specifically, $n \in \{10^2, 10^3, 10^4, 10^5, 10^6\}$. Each time, we fix a value of (β, r) such that $r > \beta$.

In the first setting, we set $(\beta, r) = (0.4, 0.9)$. The simulation results are reported in Figure 5 and Figure 6. We see that, in both models, the risks of the two procedures decrease to zero rapidly as the sample size gets larger. The BH method clearly dominates (in terms of FNP) up until $n = 10^3$, and after that the two methods behave similarly.

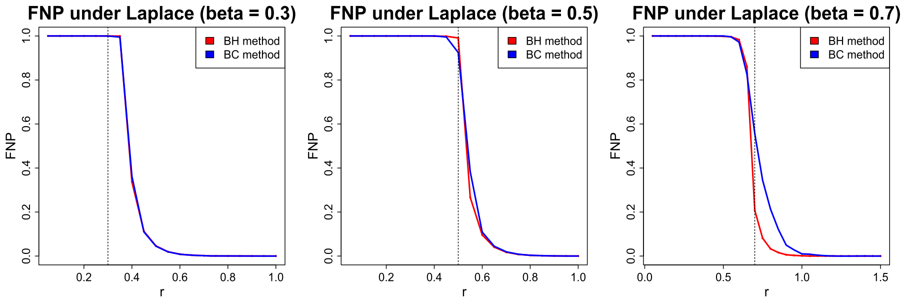

Fig 4. *Simulation results showing the FNP for the BH and BC methods under the double-exponential model in three distinct sparsity regimes. The black vertical line delineates the theoretical threshold ($r = \beta$).*

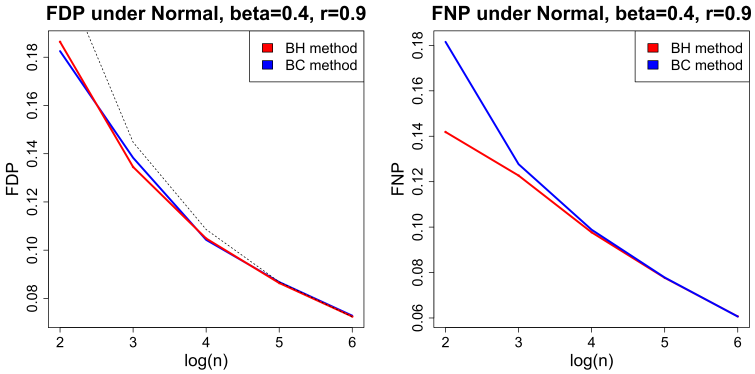

Fig 5. *FDP and FNP for the BH and BC methods under the normal model with* $(\beta, r) = (0.4, 0.9)$ *and varying sample size* n.

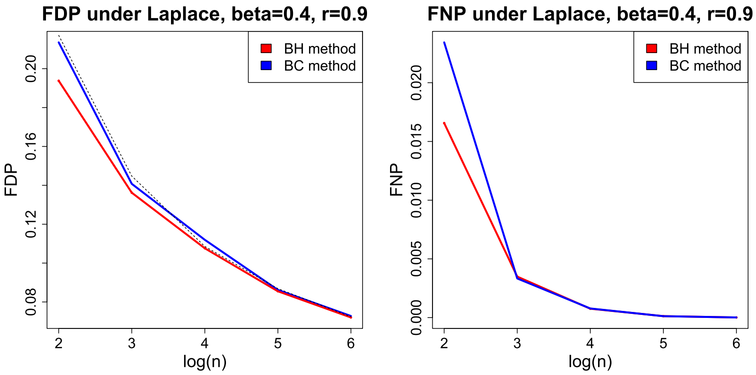

Fig 6. *FDP and FNP for the BH and BC methods under the double-exponential model with* $(\beta, r) = (0.4, 0.9)$ *and varying sample size* n.

In the second setting, we set $(\beta, r) = (0.7, 1.5)$ for normal model and $(\beta, r) = (0.7, 1.2)$ for double-exponential model. The simulation results are reported in

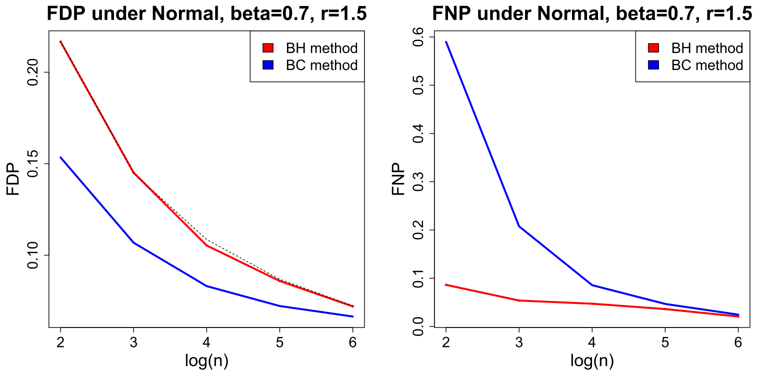

Fig 7. *FDP and FNP for the BH and BC methods under the normal model with* $(\beta, r) = (0.7, 1.5)$ *and varying sample size* n.

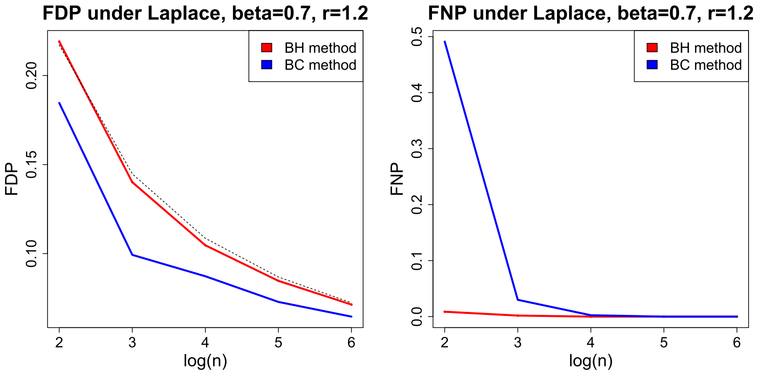

Fig 8. *FDP and FNP for the BH and BC methods under the double-exponential model with* $(\beta, r) = (0.7, 1.2)$ *and varying sample size* n.

Figure 7 and Figure 8. In this sparser regime, we can see that the BC method is much more conservative than BH method when n is relatively small. But as n gets larger, this is less pronounced. The BH method clearly dominates (in terms of FNP) up until $n = 10^3$ and past $n = 10^4$ the two methods behave similarly. The difference is much more dramatic here, in line with our findings in Section 5.1.

6. Proofs

We prove our results in this section.

6.1. Proof of Theorem 1

For $t \in \mathbb{R}$, recall that $\mathcal{R}_t = \{i : X_i \geq t\}$ and define

$$\text{the number of type I errors:} \quad \mathrm{I}(t) = |\mathcal{R}_t \setminus \mathcal{F}| \ ; \tag{15}$$

$$\text{the number of type II errors:} \quad \mathrm{II}(t) = |\mathcal{F} \setminus \mathcal{R}_t| \ . \tag{16}$$

Set $\delta = \log\log n$. We distinguish between two cases.

- When $t \le \mu + \delta$, using the fact that $t \mapsto \mathrm{I}(t)$ is non-increasing, we have

$$\textsc{fdp}(\mathcal{R}_t) = \frac{\mathrm{I}(t)}{|\mathcal{R}_t|} \ge \frac{\mathrm{I}(t)}{\mathrm{I}(t) + |\mathcal{F}|} \ge \frac{\mathrm{I}(\mu+\delta)}{\mathrm{I}(\mu+\delta) + m}. \tag{17}$$

- When $t > \mu + \delta$, using the fact that $t \mapsto \mathrm{II}(t)$ is non-decreasing, we have

$$\textsc{fnp}(\mathcal{R}_t) = \frac{\mathrm{II}(t)}{|\mathcal{F}|} \ge \frac{\mathrm{II}(\mu+\delta)}{m}. \tag{18}$$

(Recall that $m = |\mathcal{F}|$ in our model.) Hence, we conclude that for any $t \in \mathbb{R}$,

$$\textsc{fdp}(\mathcal{R}_t) + \textsc{fnp}(\mathcal{R}_t) \ge \frac{\mathrm{I}(\mu+\delta)}{\mathrm{I}(\mu+\delta) + m} \wedge \frac{\mathrm{II}(\mu+\delta)}{m}. \tag{19}$$

Consequently, to show that the oracle threshold risk has limit inferior at least 1 as n tends to infinity, by dominated convergence, it suffices to show that the RHS tends to 1 in probability, or put differently, that

$$\frac{\mathrm{I}(\mu+\delta)}{m} \to \infty \quad \text{and} \quad \frac{\mathrm{II}(\mu+\delta)}{m} \to 1, \quad \text{in probability as } n \to \infty. \tag{20}$$

On the one hand, we have $\mathrm{I}(\mu+\delta) \sim \mathrm{Bin}(n-m, \Psi(\mu+\delta))$, so that for $\mathrm{I}(\mu+\delta)/m$ to diverge to ∞ in probability it suffices that $(n-m)\Psi(\mu+\delta)/m \to \infty$. And indeed, this is the case since

$$\begin{aligned} \log\Big[(n-m)\Psi(\mu+\delta)/m\Big] &= \log(n/m) + o(1) + \log\Psi(\mu+\delta) && (21)\\ &= \log(n/n^{1-\beta}) + o(1) - \tfrac{1}{\gamma}(\mu+\delta)^\gamma(1+o(1)) && (22)\\ &= \beta\log n + o(1) - (r + o(1))\log n && (23)\\ &= (\beta - r + o(1))\log n \to \infty, && (24) \end{aligned}$$

using the fact that $m \sim n^{1-\beta}$, that Ψ is AGG with exponent γ, that $\mu + \delta \sim \mu$ with μ defined in (10), and that $r < \beta$.

On the other hand, we have $\mathrm{II}(\mu+\delta) \sim \mathrm{Bin}(m, 1-\Psi(\delta))$, so that for $\mathrm{II}(\mu+\delta)/m$ to converge to 1 in probability it suffices that $\Psi(\delta) \to 0$, which is the case since $\delta \to \infty$.

6.2. *Proof of Theorem 2*

Let Ψ denote the null survival function, assumed to be AGG with parameter $\gamma \ge 1$. Let $\hat{G}$ denote the empirical survival function

$$\hat{G}(t) = \frac{1}{n}\sum_{i\in[n]} \mathbb{I}\{X_i \ge t\}. \tag{25}$$

Let $Y_i = X_i - \mu_i$ and note that $(Y_i : i \in [n])$ are IID with distribution Ψ. Define the empirical survival functions

$$\hat{W}_{\text{true}}(y) = \frac{1}{n-m} \sum_{i \notin \mathcal{F}} \mathbb{I}\{Y_i \geq y\}, \quad \hat{W}_{\text{false}}(y) = \frac{1}{m} \sum_{i \in \mathcal{F}} \mathbb{I}\{Y_i \geq y\}, \tag{26}$$

so that

$$\hat{G}(t) = (1-\varepsilon)\hat{W}_{\text{true}}(t) + \varepsilon \hat{W}_{\text{false}}(t-\mu). \tag{27}$$

where $\varepsilon := m/n \sim n^{-\beta}$ under (9).

We need the following result to control the deviations of the empirical distributions.

Lemma 1 ([13]). *Let $Z_1, \dots, Z_k$ be IID with continuous survival function Q. Let $\hat{Q}_k$ denote their empirical survival function and define $\zeta_k = \sqrt{2 \log\log(k)/k}$ for $k \geq 3$. Then*

$$\frac{1}{\zeta_k} \max_z \frac{\hat{Q}_k(z) - Q(z)}{\sqrt{Q(z)(1-Q(z))}} \to 1, \text{ in probability as } k \to \infty. \tag{28}$$

In particular,

$$\hat{Q}_k(z) = Q(z) + O_{\mathbb{P}}(\zeta_k)\sqrt{Q(z)(1-Q(z))}, \quad \text{uniformly in } z. \tag{29}$$

Applying Lemma 1, we get

$$\hat{G}(t) = (1-\varepsilon)\big[\Psi(t) + O_{\mathbb{P}}(\zeta_n)\sqrt{\Psi(t)(1-\Psi(t))}\big] \tag{30}$$

$$+ \varepsilon\big[\Psi(t-\mu) + O_{\mathbb{P}}(\zeta_m)\sqrt{\Psi(t-\mu)(1-\Psi(t-\mu))}\big]. \tag{31}$$

From this we get

$$\hat{G}(t) = G(t) + \hat{R}(t), \tag{32}$$

where

$$G(t) := \mathbb{E}[\hat{G}(t)] = (1-\varepsilon)\Psi(t) + \varepsilon\Psi(t-\mu), \tag{33}$$

and

$$\hat{R}(t) = O_{\mathbb{P}}\big(\zeta_n \sqrt{\Psi(t)(1-\Psi(t))} + \zeta_m \varepsilon \sqrt{\Psi(t-\mu)(1-\Psi(t-\mu))}\big), \tag{34}$$

uniformly in $t \in \mathbb{R}$.

Let $\iota = \iota_{\text{BH}}$ be defined as in (11). We have $\hat{G}(X_{(i)}) = i/n$, so that $X_{(i)} \geq \Psi^{-1}(q\hat{G}(X_{(i)}))$ for $i \leq \iota$ and $X_{(i)} < \Psi^{-1}(q\hat{G}(X_{(i)}))$ for $i > \iota$. Based on that, and the fact that $\hat{G}$ is constant between two consecutive X_i's, we have that there is $\tau \in (X_{(\iota+1)}, X_{(\iota)}]$ such that

$$\tau = \min\big\{t : t \geq \Psi^{-1}(q\hat{G}(t))\big\} = \min\big\{t : t = \Psi^{-1}(q\hat{G}(t))\big\}. \tag{35}$$

Note that the BH procedure coincides with $\mathcal{R}_\tau$, the threshold method with threshold τ. In particular,

$$\text{FNP}(\mathcal{R}_\tau) = 1 - \hat{F}(\tau), \quad \hat{F}(t) := \frac{1}{m} \sum_{i \in \mathcal{F}} \mathbb{I}\{X_i \geq t\}, \tag{36}$$

so that it suffices to show that $\hat{F}(\tau) \to 1$ in probability. As above, by Lemma 1,

$$\hat{F}(t) = \hat{W}_{\text{false}}(t-\mu) = \Psi(t-\mu) + O_{\mathbb{P}}(\zeta_m)\sqrt{\Psi(t-\mu)(1-\Psi(t-\mu))}, \tag{37}$$

and in particular $\hat{F}(\tau) = \Psi(\tau-\mu) + o_{\mathbb{P}}(1)$, so it suffices to show that $\tau - \mu \to -\infty$ in probability.

Since $r > \beta$ and $\beta < 1$, we may take a real number $r_* \in (\beta, r \wedge 1)$. Define $t_* = (\gamma r_* \log n)^{1/\gamma}$. Since $t_* - \mu \to -\infty$, it suffices to show that $\tau \le t_*$ with probability tending to 1. We have

$$G(t_*) = (1-\varepsilon)\Psi(t_*) + \varepsilon \Psi(t_* - \mu). \tag{38}$$

The first term is $\sim \Psi(t_*)$, with

$$\Psi(t_*) = n^{-r_* + o(1)}, \tag{39}$$

by Definition 1, which says that $\log \Psi(t) \sim -t^\gamma/\gamma$ as $t \to \infty$. The second term is $\sim n^{-\beta}$ by (9) and the fact that $\Psi(t_* - \mu) \to 1$ since, again, $t_* - \mu \to -\infty$. Together, we obtain $G(t_*) \sim n^{-\beta}$, using also the fact that $r_* > \beta$. In addition, by (34) we have

$$\hat{R}(t_*) = O_{\mathbb{P}}\big(\zeta_n \sqrt{\Psi(t_*)}\big) + o_{\mathbb{P}}(\varepsilon) = o_{\mathbb{P}}(n^{-\beta}), \tag{40}$$

since $\zeta_n \sqrt{\Psi(t_*)} = n^{-\frac{1}{2}(r_*+1)+o(1)}$ (any poly-logarithmic factor was absorbed in $n^{o(1)}$), again by (39), and $\beta < r_* < 1$. Hence, applying (32), we obtain

$$\hat{G}(t_*) = G(t_*) + \hat{R}(t_*) \sim_{\mathbb{P}} G(t_*) \sim n^{-\beta}. \tag{41}$$

Together with (39), and using by (12), we have

$$\hat{G}(t_*)/\Psi(t_*) = n^{(r_*-\beta)+o_{\mathbb{P}}(1)} \gg 1/q. \tag{42}$$

This, together with (35), implies that $\tau \le t_*$ with probability tending to 1, hence $\text{FNP}(\mathcal{R}_\tau) \to 0$ in probability. By Remark 6, we have $\text{FNR}(\mathcal{R}_\tau) \to 0$ as $n \to \infty$.

6.3. Proof of Theorem 3

The proof borrows a number of arguments from Section 6.2. We use the same notation and assume as before that the X_i's are distinct. We order the absolute values of statistic $|X|$ in decreasing order, meaning that $|X|_{(1)} \ge \cdots \ge |X|_{(n)}$. Recall that Ψ is now symmetric about 0.

Define the threshold

$$\tau = \inf\big\{t : \widehat{\text{FDP}}(t) \le q\big\}. \tag{43}$$

The difference with τ_{BC} in (13) is that the range is not limited to $|\mathbf{X}|$. It can be seen that $\tau = |X|_{(\iota_{\text{BC}}+1)}$ if $\iota_{\text{BC}} < n$ and $\tau = 0$ if $\iota_{\text{BC}} = n$. This, in particular, implies

$$\text{FNP}(\mathcal{R}_\tau) \le \text{FNP}(\mathcal{R}_{\tau_{\text{BC}}}) \le \text{FNP}(\mathcal{R}_\tau) + \tfrac{1}{m}. \tag{44}$$

Since in our model $m \to \infty$, it suffices to show that $\text{FNP}(\mathcal{R}_\tau) \to 0$ in probability. As before, (36) holds true, so it suffices to show that $\hat{F}(\tau) \to 1$ in probability. For that, we saw earlier that it suffices to show that $\tau \le t_*$ with probability tending to 1.

We have

$$\widehat{\text{FDP}}(t_*) = \frac{1 + n(1 - \hat{G}(-t_*))}{1 \vee n\hat{G}(t_*)}. \tag{45}$$

We already saw that $\hat{G}(t_*) \sim n^{-\beta}$, so the denominator above is $\sim n^{1-\beta}$ as $n \to \infty$. For the numerator, by (32), we have

$$1 - \hat{G}(-t_*) = 1 - G(-t_*) - \hat{R}(-t_*). \tag{46}$$

By (33),

$$\begin{aligned} 1 - G(-t_*) &= (1-\varepsilon)(1 - \Psi(-t_*)) + \varepsilon(1 - \Psi(-t_* - \mu)) && (47)\\ &= (1-\varepsilon)\Psi(t_*) + \varepsilon\Psi(t_* + \mu) \quad [\text{by symmetry of } \Psi] && (48)\\ &\sim \Psi(t_*) = n^{-r_* + o(1)}. \quad [\text{by } (39)] && (49) \end{aligned}$$

By (34),

$$\begin{aligned} \hat{R}(-t_*) &= O_{\mathbb{P}}\big(\zeta_n\sqrt{1 - \Psi(-t_*)} + \zeta_m\varepsilon\sqrt{1 - \Psi(-t_* - \mu)}\big) && (50)\\ &= O_{\mathbb{P}}(\zeta_n\sqrt{\Psi(t_*)} + \zeta_m\varepsilon\sqrt{\Psi(t_* + \mu)}) \quad [\text{by symmetry of } \Psi)] && (51)\\ &= O_{\mathbb{P}}(n^{-\frac{1}{2}(r_*+1)+o(1)} + o(n^{-\frac{1}{2}(r_*+\beta+1)+o(1)})) \quad [\text{by } (39)] && (52)\\ &= O_{\mathbb{P}}(n^{-\frac{1}{2}(r_*+1)+o(1)}). && (53) \end{aligned}$$

(Again, any poly-logarithmic factor was absorbed in $n^{o(1)}$.) Combined with the fact that $r_* < 1$, we get $1 - \hat{G}(-t_*) \sim n^{-r_*+o(1)}$, and therefore

$$\widehat{\text{FDP}}(t_*) = \frac{n^{1-r_*+o(1)}}{n^{1-\beta}} = n^{\beta - r_* + o(1)} \ll q. \quad [\text{by } (12) \text{ and } \beta < r_\star] \tag{54}$$

Hence, $\widehat{\text{FDP}}(t_*) \le q$ with probability tending to 1, and when this is the case, $\tau \le t_*$, by definition of τ above. This also implies $\text{FNP}(\mathcal{R}_\tau) \to 0$ in probability. By Remark 6, we have $\text{FNR}(\mathcal{R}_\tau) \to 0$ as $n \to \infty$.

Acknowledgments

This work was partially supported by grants from the US National Science Foundation (DMS 1223137) and the US Office of Naval Research (N00014-13-1-0257). We would like to thank Chao Gao for providing the reference [29], as well as two anonymous referees for providing important pointers to the literature that we had originally missed, in addition to other useful comments.

References

[1] Abramovich, F., Y. Benjamini, D. L. Donoho, and I. M. Johnstone (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics 34*(2), 584–653. MR2281879

[2] Arias-Castro, E. and M. Wang (2016). Distribution-free tests for sparse heterogeneous mixtures. *TEST*, 1–24. arXiv preprint arXiv:1308.0346. MR3613606

[3] Arlot, S., G. Blanchard, and E. Roquain (2010). Some non-asymptotic results on resampling in high dimension, II: Multiple tests. *The Annals of Statistics 38*(1), 83–99. MR2589317

[4] Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(1), 289–300. MR1325392

[5] Bogdan, M., A. Chakrabarti, F. Frommlet, and J. K. Ghosh (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 1551–1579. MR2850212

[6] Butucea, C., N. A. Stepanova, and A. B. Tsybakov (2015). Variable selection with hamming loss. *arXiv preprint arXiv:1512.01832*.

[7] Cai, T. T. and J. Jin (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. *The Annals of Statistics 38*(1), 100–145. MR2589318

[8] Dickhaus, T. (2014). *Simultaneous statistical inference*. Springer. MR3184277

[9] Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics 32*(3), 962–994. MR2065195

[10] Dudoit, S. and M. J. van der Laan (2007). *Multiple testing procedures with applications to genomics*. Springer Science & Business Media. MR2373771

[11] Durot, C. and Y. Rozenholc (2006). An adaptive test for zero mean. *Mathematical Methods of Statistics 15*(1), 26–60. MR2225429

[12] Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association 99*(465), 96–104. MR2054289

[13] Eicker, F. (1979). The asymptotic distribution of the suprema of the standardized empirical processes. *The Annals of Statistics 7*(1), 116–138. MR0515688

[14] Foygel-Barber, R. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics 43*(5), 2055–2085. MR3375876

[15] Ge, Y., S. Dudoit, and T. P. Speed (2003). Resampling-based multiple testing for microarray data analysis. *Test 12*(1), 1–77. MR1993286

[16] Genovese, C. and L. Wasserman (2002). Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64*(3), 499–517. MR1924303

[17] Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc. MR0758442

[18] Ingster, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics 6*(1), 47–69. MR1456646

[19] Ingster, Y. I. and I. A. Suslina (2003). *Nonparametric goodness-of-fit testing under Gaussian models*, Volume 169 of *Lecture Notes in Statistics*. New York: Springer-Verlag. MR1991446

[20] Ji, P., J. Jin, et al. (2012). Ups delivers optimal phase diagram in high-dimensional variable selection. *The Annals of Statistics 40*(1), 73–103. MR3013180

[21] Jin, J. and T. T. Cai (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association 102*(478), 495–506. MR2325113

[22] Jin, J. and T. Ke (2014). Rare and weak effects in large-scale inference: methods and phase diagrams. *arXiv preprint arXiv:1410.4578*. MR3468343

[23] Meinshausen, N., M. H. Maathuis, P. Bühlmann, et al. (2011). Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics 39*(6), 3369–3391. MR3012412

[24] Neuvial, P. and E. Roquain (2012). On false discovery rate thresholding for classification under sparsity. *The Annals of Statistics 40*(5), 2572–2600. MR3097613

[25] Pollard, K. S. and M. J. van der Laan (2004). Choice of a null distribution in resampling-based multiple testing. *Journal of Statistical Planning and Inference 125*(1), 85–100. MR2086890

[26] Romano, J. P. and M. Wolf (2007). Control of generalized error rates in multiple testing. *The Annals of Statistics*, 1378–1408. MR2351090

[27] Roquain, E. (2011). Type i error rate control in multiple testing: a survey with proofs. *Journal de la Société Française de Statistique 152*(2), 3–38. MR2821220

[28] Storey, J. D. (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(3), 347–368. MR2323757

[29] Sun, W. and T. T. Cai (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association 102*(479), 901–912. MR2411657

[30] Westfall, P. H. and S. S. Young (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Volume 279. John Wiley & Sons.

[31] Yekutieli, D. and Y. Benjamini (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference 82*(1), 171–196. MR1736442